45

38-10(15493)D

#11
CO
11/12/02

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of David F. Bush *et al.*	Art Unit: 1634
Serial No: 09/803,736	Examiner: Einsmann, Juliet Caroline
Filed : 03/12/2001	
For: Plant Polymorphic Markers and Uses Thereof	

RESPONSE TO RESTRICTION REQUIREMENT AND
REQUIREMENT FOR INFORMATION UNDER 37 CFR 1.105

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

Responsive to the further restriction requirement set forth in the Patent Office communication dated October 8, 2002 and setting forth a one month period for response, Applicants provide the following.

Applicants traverse the further undue division of their invention, e.g. which effectively requires applicant to file at least 50 applications to cover the full scope of their invention of claim 18 which is characterized by 100 identified polymorphic markers listed in a Markush Group. Applicants submit that such a restriction is an improper attempt by the PTO to rewrite claim 18 to cover another invention (which is not what applicant wishes to claim) and thus is effectively an improper rejection of claim 18 under 35 U.S.C. 121. Imposing this type of effective rejection for a Markush claim is improper as a matter of law. *In re Weber*, 580 F.2d. 455,459, 198 U.S.P.Q. 328,332 (C.C.P.A. 1978) (holding that a rejection [of a Markush claim] under §121 violates the basic right of the applicant to claim his invention as he chooses"). Applicants reserve the right to appeal this rejection. Applicants find the rationale that a burden is created on USPTO computer resources running a larger search to be an unpersuasive attempt to

justify the improper examination. Technical incompetence in the leading patent examining institution will not be tolerated. The further recitation of burden on a examiner having to "wade through reams and reams of results in order to evaluate each polymorphism with regard to prior art" is likewise dismissed as pitiful excuse for denying examination of Applicants' invention. Wading through reams would be an archaic method when the PTO could easily construct a search algorithm to display results in order of novelty, e.g. ranked by percent identity, which would allow examination for the few sequences at the novelty interface.

However, to comply with the restriction requirement to select a set of two polymorphisms to which examination of claims 18 and 19 will be limited and advance the prosecution of this application, Applicants hereby provisionally elect SEQ ID NOs: 466799 and 471736.

Responsive to the Requirement for Information under 37 CFR 1.105 in the above referenced communication from the Patent Office, Applicants provide the following:

Re paragraph 2 of the Requirement for Information.

Copies of two publications co-authored by one or more inventors named on the present application are enclosed.

> Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature 408*:796-814 (December 14, 2000). S. Rounsley *et al.* are authors of a section of this paper entitled "Comparative analysis of the genomes of *A. thaliana* accessions" which appears at page 801 of the publication. No portion of the marker dataset was disclosed in conjunction with this publication, although a reference to the TAIR website was provided.

Jander *et al.* (2002) Arabidopsis Map-Based Cloning in the Post-Genome Era, *Plant Physiology 129*:440-450. A small portion of the marker dataset was disclosed in this publication.

Re paragraph 3 of the Requirement for Information.

A printout of pages available at the TAIR website is attached. The pages provide information on the date of initial availability of the Cereon Arabidopsis marker dataset and the contents and dates of availability of updates to the dataset. Applicants note that registration is not required for access to this summary information available at the website www.arabidopsis.org/Cereon/help.html. Furthermore, if the Examiner wishes to personally access the marker information at the above website, the below undersigned representative of Applicants' would be happy to provide a user name and password to allow the Examiner to access the information. In response to parts a. and b. of this request, Applicants provide the following summary of the data releases and their relation to the present application and its priority documents. Importantly, Applicants note that any information in the present application that was released on the TAIR website, was released after the information was filed in either the present application or one of its priority applications..

Release 1: Released May 3, 2000

25,274 SNPs
14,041 InDels

To the best of Applicants' knowledge, the 25,274 SNPs in Release 1 correspond to the 25,274 SNPs presented in Applicants' priority application filed March 29, 2000 (serial number 09/534,859).

The 14,041 InDels in Release 1 are all, to the best of Applicants' knowledge, present in Applicants' priority application filed March 29, 2000 (serial number 09/534,859). The priority application also contained additional InDels larger than 100 bp that were excluded from the Release 1 dataset on the initial Release date as indicated on the enclosed information from the TAIR website.

Release 2: Released November 16, 2000

2843 SNPs
1633 InDels

To the best of Applicants' knowledge, the 2843 SNPs in Release 2 correspond to the 2843 SNPs presented in Applicants' priority application filed October 20, 2000 (serial number 09/692,412).

The 1633 InDels in Release 2 are all, to the best of Applicants' knowledge, present in Applicants' priority application filed October 20, 2000 (serial number 09/692,412). This application also contained additional InDels larger than 100 bp that were excluded from the Release 2 dataset at it's initial release date as indicated on the enclosed information from the TAIR website.

Release 3: Released March 21, 2001

9226 SNPs
2905 InDels
42 Large InDels

Applicants note that although the release notes indicate the release of 9227 SNPs and 43 large InDels, the actual numbers were 9226 and 42, as determined by downloading the marker set from the TAIR site.

To the best of Applicants' knowledge, the 9226 SNPs in Release 3 are all present in the set of 37,343 SNPs in the present application filed on March 12, 2001. Applicants note that the marker set in the present application is cumulative and contains the SNPs from all three TAIR releases.

The 2905 InDels and 42 large InDels in Release 3 are all, to the best of Applicants' knowledge, present in the present application filed on March 12, 2001.

In response to part c. of this request, Applicants note that information on the data format requested by the Examiner is provided at page 2 of the TAIR website printout provided herein. Applicants note that polymorphism locations are provided by chromosome and with reference to their location in the public BACs referenced in the tables. Oligonucleotides for amplification are not provided, but 20 bp of sequence directly to the left and right of the polymorphism is provided.

Re paragraph 4 of the Requirement for Information.

The Examiner has also requested information on the location of the elected markers in Table A of the application. Applicants provide herewith the printed pages of Table A that contain the information on markers 466799 and 471736. In Applicants' records, the information for marker 466799 is found at page 159/208 of Table A, and the information for marker 471736 is found at page 75/208 of Table A. Applicants note that a paper copy of the table as submitted with the application was not retained, and it is not known if the above listed page numbers correspond with the page numbers of the record copy of the application. For the Examiner's convenience, in addition to the printed pages, we have extracted the information on the elected markers from Table A and provide this information on a separate sheet in table format, including the table headings.

38-10(15493)D

CHANGE OF CORRESPONDENCE ADDRESS

Please direct all future correspondence to:

Gail Wuelner, E2NA
Monsanto Company
800 N. Lindbergh Blvd.
St.Louis, MO 63167

Please direct future telephone calls to :
Thomas E. Kelley
Registration No. 29,938
(860) 572-5274

Respectfully submitted,

Thomas E. Kelley
Registration No. 29,938
Applicant's Attorney
Phone: (860) 572-5274

CEREON ARABIDOPSIS POLYMORPHISM COLLECTION HELP

Release 1 README | Release 2 README | Release 3 README | Column Headings | Notes and Usage

The accompanying tables contains potential nucleotide polymorphisms between Columbia and Landsberg erecta ecotypes. The data were generated by Cereon Genomics using all the sequenced BACs up to the point of each release. The clones listed in the table are ordered by chromosome and their location on a chromosome.

The data were kindly submitted by Dr. Steve Rounsley. All questions referring to the data should be sent to: athal@cereon.com

Release 1: Released May 3, 2000

SNPs and INDELs
There are 39315 entries (from 981 BACs). 25,274 SNPs (Single Nucleotide Polymorphism) and 14,041 INDELs (small insertions/deletions).
Large INDELs NEW
632 INDELs > 100bp found in 341 of the 980 BACs used for Release 1.
Size range: Min 101bp Max 38Kb
This collection of INDELs are larger than 100bp and were omitted from the original polymorphism release due to an increased level of false positives contained within. Repetitive sequence is the major cause for such false positives, and the larger the gap between two matches, the more likely one of the matches is due to a match against a repeated region.
The data is being provided with these caveats so please use this with caution. There are however true positives in this dataset, particularly the INDELs at the lower end of the size distribution, and many people had requested access to these for the purposes of specific studies such as transposon analysis.

Release 2: Released on November 16, 2000

SNPs and INDELs NEW
Release 2 contains 4476 predicted polymorphisms from 124 BAC clones that have been sequenced by the AGI. 1,633 INDELs and 2843 SNPs were found. This is an 11% increase over the previous collection of approximately 39,000.
The remainder of the Columbia BACs are currently being processed and there will be a final update of the Cereon collection shortly after this is completed.
Large INDELs NEW
72 Large INDELs found in 43 of the 124 BACs used for Release 2.
Also provided here are INDELs greater than 100 bp. These should be treated with caution, as they are more likely to be the result of artifacts of the analysis method., However, many will still be true insertions/deletions and may therefore be of interest for certain kinds of analysis.

Release 3: Released on March 21, 2001

SNPs and INDELs NEW
Release 3 contains 12175 predicted polymorphisms from 396 BAC clones that have been sequenced by the AGI. 2905 INDELs, 9227 SNPs, and 43 Large INDELs were found. This is an 27% increase over the previous collection of approximately 45,000. The total number of polymorphisms now is 56,670.

Column Headings:
1. SNP_Name: CER(Cereon)+Cereon's internal ID number e.g. CER454879
2. Chromosome: 1 thru 5
3. BAC_Name: in standard AGI format, ordered by position in chromosome
4. BAC_Accession: GenBank/EMBL/DDBJ accession number
5. BAC_Length: in bp
6. Left_Coor*: The coordinate of the base to the left of the polymorphic location. See below.
7. Right_Coor*: The coordinate of the base to the left of the polymorphic location. See below.
8. SNP_Type: SNP (Single Nucleotide Polymorphism) or IND (Insertion/Deletion)
9. IND_Size: size of insertion or deletion in Columbia. Col/Ler. e.g. -4/4 means a 4bp deletion in Columbia and 4/-4 means a 4bp insertion in Col. Left blank for SNPs.
10. SNP_Base: changed nucleotide. Col/Ler e.g. A/T means A in Columbia and T in Landsberg. Left blank for INDELs
11. Left_Flank: 20 bp directly to the left of the polymorphic location
12. Right_Flank: 20bp directly to the right of the polymorphic location. See Note below.
13. Restriction_sites (Col): restriction sites in Col from the SNP/IND
14. Restriction_sites (Ler): restriction sites in Ler from the SNP/IND

*Coordinates:
For a SNP - the two coordinates flank the polymorphic base.
For an insertion in Columbia, the two coordinates flank the inserted sequence.
For a deletion in Columbia, the deletion occurs between the two coordinates listed.

NOTE:
1. 20mers are provided for locating the correct coordinates just in case the BAC sequence changes. Also note that the 20mers were supposed to be directly flanking the changed nucleotide, but the right flanking sequence may be off by one base for SNPs - but not for INDs.

Please note that the coordinates used in the datafiles refer to the originally submitted BAC sequence. Many BAC sequences at GenBank have been edited by the AGI groups in order to produce finished chromosome records. This involves removing overlapping regions, and flipping some clones in order to produce a consistent direction along the chromosome. In addition, AGI groups may make alterations at any time to the submitted sequence in order to correct errors. This can also cause the original coordinates to be inaccurate.

In order to access the original BAC sequence, you need to use the link provided in the current GenBank record. The link will look something like this:

"COMMENT: On Dec 16, 1999 this sequence version replaced gi:5729683"

USAGE:
The flanking sequence provided in the Cereon data files attempts to provide an alternative way to locate the polymorphism. The 20mers can be used to BLAST against the Arabidopsis genome to identify the specified location. There are some caveats to keep in mind when doing this:

- 1. This sequence should help find the appropriate location in the BAC of interest. It is not necessarily unique to the genome. It may also match other BACs in the genome, but these are not important for locating the polymorphism.
- 2. If the 20mer matches more than once in the BAC of interest, try using the other 20mer as well and combining the results. You can also use TAIR's PatMatch, which allows you to put in the polymorphic sequence as well as its approximate length in between the two 20mer set.
- 3. If the 20mer does not find a match in the BAC of interest, it could be that the editing mentioned above may have moved this location to a neighboring BAC. In this case, check your search results against the neighboring BACs.
- 4. If it still does not match, beware that using the default BLAST parameters does not always work well with such a small query sequence. Several things can increase your chances of finding a match in the BAC sequence of interest.
    - A. Use a smaller database. An example would be a species specific collection at NCBI, or the TAIR BLAST server selecting only Arabidopsis genomic sequences > 10kb

- B. Do not filter for low complexity.
- C. Increase the mismatch penalty to -8. This should force identical matches.

- 5. If multiple hits to the same BAC occur - do not panic. Remember, many INDELs are caused by a different copy number of a direct repeat. The flanking sequence may therefore hit multiple places. The best bet here is to pick primers several hundred bases either side of this general region.

2	AC006837	87584	F23H14468703	86107	86109	SNP	1		C/A
2	AC006837	87584	F23H14469865	12879	12920	IND	2	40/-40	
2	AC006837	87584	F23H14469866	13133	16333	IND	2	3199/-3199	
2	AC006837	87584	F23H14469867	19400	19401	IND	2	-7/7	
2	AC006837	87584	F23H14469868	22592	22629	IND	2	36/-36	
2	AC006837	87584	F23H14469869	22932	28384	IND	2	5451/-5451	
2	AC006837	87584	F23H14469870	24106	24107	IND	2	-14/14	
2	AC006837	87584	F23H14469871	26527	26528	IND	2	-4/4	
2	AC006837	87584	F23H14469872	28142	42196	IND	2	14053/-14053	
2	AC006837	87584	F23H14469873	47873	47874	IND	2	-4/4	
2	AC006837	87584	F23H14469874	61687	64076	IND	2	2388/-2388	
2	AC006837	87584	F23H14469875	62060	62061	IND	2	-4/4	
2	AC006837	87584	F23H14469876	62720	62751	IND	2	30/-30	
2	AC006837	87584	F23H14469877	82080	82090	IND	2	9/-9	
2	AC007730	94503	T5M2	471731	10774	10776	SNP	1	G/A
2	AC007730	94503	T5M2	471732	10426	10428	SNP	1	A/C
2	AC007730	94503	T5M2	471733	10465	10467	SNP	1	A/C
2	AC007730	94503	T5M2	471734	10332	10334	SNP	1	C/G
2	AC007730	94503	T5M2	471735	10360	10362	SNP	1	C/T
2	AC007730	94503	T5M2	471736	10623	10625	SNP	1	A/T
2	AC007584	82189	MJB20	472107	54110	54112	SNP	1	T/A
2	AC007584	82189	MJB20	472164	71962	71964	SNP	1	A/C

3	AL132972	77350	T25B15468164	53542	53544	SNP	1		T/A
3	AL132972	77350	T25B15468165	55866	55868	SNP	1		T/C
3	AL132972	77350	T25B15468166	56131	56133	SNP	1		C/G
3	AL132972	77350	T25B15468167	56111	56113	SNP	1		G/T
3	AL132972	77350	T25B15468374	28406	28408	SNP	1		G/T
3	AL132972	77350	T25B15468555	18172	18174	SNP	1		A/C
3	AL132972	77350	T25B15469018	65343	65345	SNP	1		T/A
3	AL132972	77350	T25B15469019	65586	65588	SNP	1		A/T
3	AL132972	77350	T25B15470856	42246	42265	IND	2	18/-18	
3	AL132972	77350	T25B15471540	18231	18232	IND	1	-1/1	
3	AL132972	77350	T25B15471541	65396	65397	IND	1	-1/1	
3	AL132972	77350	T25B15471542	65453	65455	IND	1	1/-1	
3	AL132969	97798	F8J2	466797	12607	12609	SNP	1	T/A
3	AL132969	97798	F8J2	466798	13819	13821	SNP	1	A/G
3	AL132969	97798	F8J2	466799	13671	13673	SNP	1	C/T
3	AL132969	97798	F8J2	466829	17228	17230	SNP	1	C/T
3	AL132969	97798	F8J2	467141	67333	67335	SNP	1	T/C
3	AL132969	97798	F8J2	467142	66591	66593	SNP	1	C/G
3	AL132969	97798	F8J2	467159	207	209	SNP	1	G/A
3	AL132969	97798	F8J2	467160	486	488	SNP	1	T/C
3	AL132969	97798	F8J2	467161	478	480	SNP	1	A/C

Information on Elected Markers Extracted from Table A

Chromosome	Accession	BAC Length	BAC Name	MarkerName	Left	Right	Type	Method	Indel Size Columbia/Landsberg	SNP Base Columbia/Landsberg
3	AL132969	97798	F8J2	466799	13671	13673	SNP	1		G/T
2	AC007730	94503	T5M2	471736	10623	10625	SNP	1		A/T

articles

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*

The Arabidopsis Genome Initiative*

**Authorship of this paper should be cited as 'The Arabidopsis Genome Iniative'. A full list of contributors appears at the end of this paper*

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions. Here we report the analysis of the genomic sequence of *Arabidopsis*. The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *Caenorhabditis elegans*— the other sequenced multicellular eukaryotes. *Arabidopsis* has many families of new proteins but also lacks several common protein families, indicating that the sets of common proteins have undergone differential expansion and contraction in the three multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

The plant and animal kingdoms evolved independently from unicellular eukaryotes and represent highly contrasting life forms. The genome sequences of *C. elegans*[1] and *Drosophila*[2] reveal that metazoans share a great deal of genetic information required for developmental and physiological processes, but these genome sequences represent a limited survey of multicellular organisms. Flowering plants have unique organizational and physiological properties in addition to ancestral features conserved between plants and animals. The genome sequence of a plant provides a means for understanding the genetic basis of differences between plants and other eukaryotes, and provides the foundation for detailed functional characterization of plant genes.

Arabidopsis thaliana has many advantages for genome analysis, including a short generation time, small size, large number of offspring, and a relatively small nuclear genome. These advantages promoted the growth of a scientific community that has investigated the biological processes of *Arabidopsis* and has characterized many genes[3]. To support these activities, an international collaboration (the Arabidopsis Genome Initiative, AGI) began sequencing the genome in 1996. The sequences of chromosomes 2 and 4 have been reported[4,5], and the accompanying Letters describe the sequences of chromosomes 1 (ref. 6), 3 (ref. 7) and 5 (ref. 8).

Here we report analysis of the completed *Arabidopsis* genome sequence, including annotation of predicted genes and assignment of functional categories. We also describe chromosome dynamics and architecture, the distribution of transposable elements and other repeats, the extent of lateral gene transfer from organelles, and the comparison of the genome sequence and structure to that of other *Arabidopsis* accessions (distinctive lines maintained by single-seed descent) and plant species. This report is the summation of work by experts interested in many biological processes selected to illuminate plant-specific functions including defence, photomorphogenesis, gene regulation, development, metabolism, transport and DNA repair.

The identification of many new members of receptor families, cellular components for plant-specific functions, genes of bacterial origin whose functions are now integrated with typical eukaryotic components, independent evolution of several families of transcription factors, and suggestions of as yet uncharacterized metabolic pathways are a few more highlights of this work. The implications of these discoveries are not only relevant for plant biologists, but will also affect agricultural science, evolutionary biology, bioinformatics, combinatorial chemistry, functional and comparative genomics, and molecular medicine.

Overview of sequencing strategy

We used large-insert bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) libraries[9–12] as the primary substrates for sequencing. Early stages of genome sequencing used 79 cosmid clones. Physical maps of the genome of accession Columbia were assembled by restriction fragment 'fingerprint' analysis of BAC clones[13], by hybridization[14] or polymerase chain reaction (PCR)[15] of sequence-tagged sites and by hybridization and Southern blotting[16]. The resulting maps were integrated (http://nucleus/cshl.org/arabmaps/) with the genetic map and provided a foundation for assembling sets of contigs into sequence-ready tiling paths. End sequence (http://www.tigr.org/tdb/at/abe/bac_end_search.html) of 47,788 BAC clones was used to extend contigs from BACS anchored by marker content and to integrate contigs.

Ten contigs representing the chromosome arms and centromeric heterochromatin were assembled from 1,569 BAC, TAC, cosmid and P1 clones (average insert size 100 kilobases (kb)). Twenty-two PCR products were amplified directly from genomic DNA and sequenced to link regions not covered by cloned DNA or to optimize the minimal tiling path. Telomere sequence was obtained from specific yeast artificial chromosome (YAC) and phage clones, and from inverse polymerase chain reaction (IPCR) products derived from genomic DNA. Clone fingerprints, together with BAC end sequences, were generally adequate for selection of clones for sequencing over most of the genome. In the centromeric regions, these physical mapping methods were supplemented with genetic mapping to identify contig positions and orientation[17].

Selected clones were sequenced on both strands and assembled using standard techniques. Comparison of independently derived sequence of overlapping regions and independent reassembly sequenced clones revealed accuracy rates between 99.99 and 99.999%. Over half of the sequence differences were between genomic and BAC clone sequence. All available sequenced genetic markers were integrated into sequence assemblies to verify sequence contigs[4–8]. The total length of sequenced regions, which extend from either the telomeres or ribosomal DNA repeats to the 180-base-pair

(bp) centromeric repeats, is 115,409,949 bp (Table 1). Estimates of the unsequenced centromeric and rDNA repeat regions measure roughly 10 megabases (Mb), yielding a genome size of about 125 Mb, in the range of the 50–150 Mb haploid content estimated by different methods[18]. In general, features such as gene density, expression levels and repeat distribution are very consistent across the five chromosomes (Fig. 1), and these are described in detail in reports on individual chromosomes[4–8] and in the analysis of centromere, telomere and rDNA sequences.

We used tRNAscan-SE 1.21 (ref. 19) and manual inspection to identify 589 cytoplasmic transfer RNAs, 27 organelle-derived tRNAs and 13 pseudogenes—more than in any other genome sequenced to date. All 46 tRNA families needed to decode all possible 61 codons were found, defining the completeness of the functional set. Several highly amplified families of tRNAs were found on the same strand[6]; excluding these, each amino acid is decoded by 10–41 tRNAs.

The spliceosomal RNAs (U1, U2, U4, U5, U6) have all been experimentally identified in *Arabidopsis*. The previously identified sequences for all RNAs were found in the genome, except for U5 where the most similar counterpart was 92% identical. Between 10 and 16 copies of each small nuclear RNA (snRNA) were found across all chromosomes, dispersed as singletons or in small groups.

The small nucleolar RNAs (snoRNAs) consist of two subfamilies, the C/D box snoRNAs, which includes 36 *Arabidopsis* genes, and the H/ACA box snoRNAs, for which no members have been identified in *Arabidopsis*. U3 is the most numerous of the C/D box snoRNAs, with eight copies found in the genome. We identified forty-five additional C/D box snoRNAs using software (www.rna.wustl.edu/snoRNAdb/) that detects snoRNAs that guide ribose methylation of ribosomal RNA.

A combination of algorithms, all optimized with parameters based on known *Arabidopsis* gene structures, was used to define gene structure. We used similarities to known protein and expressed sequence tag (EST) sequence to refine gene models. Eighty per cent of the gene structures predicted by the three centres involved were completely consistent, 93% of ESTs matched gene models, and less than 1% of ESTs matched predicted non-coding regions, indicating

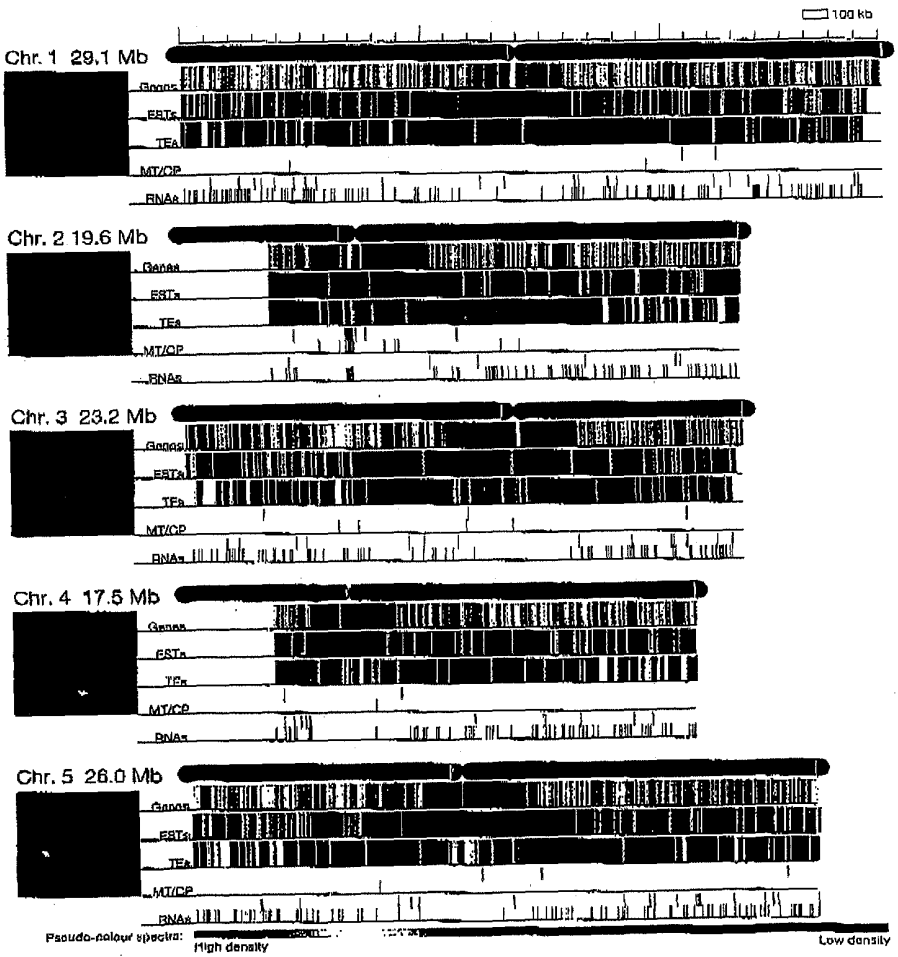

Figure 1 Representation of the *Arabidopsis* chromosomes. Each chromosome is represented as a coloured bar. Sequenced portions are red, telomeric and centromeric regions are light blue, heterochromatic knobs are shown black and the rDNA repeat regions are magenta. The unsequenced telomeres 2N and 4N are depicted with dashed lines. Telomeres are not drawn to scale. Images of DAPI-stained chromosomes were kindly supplied by P. Fransz. The frequency of features was given pseudo-colour assignments, from red (high density) to deep blue (low density). Gene density ('Genes') ranged from 38 per 100 kb to 1 gene per 100 kb; expressed sequence tag matches ('ESTs') ranged from more than 200 per 100 kb to 1 per 100 kb. Transposable element densities ('TEs') ranged from 33 per 100 kb to 1 per 100 kb. Mitochondrial and chloroplast insertions ('MT/CP') were assigned black and green tick marks, respectively. Transfer RNAs and small nucleolar RNAs ('RNAs') were assigned black and red ticks marks, respectively.

that most potential genes were identified. The sensitivity and selectivity of the gene prediction software used in this report has been comprehensively and independently assessed[20].

The 25,498 genes predicted (Table 1) is the largest gene set published to date: *C. elegans*[1] has 19,099 genes and *Drosophila*[2] 13,601 genes. *Arabidopsis* and *C. elegans* have similar gene density, whereas *Drosophila* has a lower gene density; *Arabidopsis* also has a significantly greater extent of tandem gene duplications and segmental duplications, which may account for its larger gene set.

The rDNA repeat regions on chromosomes 2 and 4 were not sequenced because of their known repetitive structure and content. The centromeric regions are not completely sequenced owing to large blocks of monotonic repeats such as 5S rDNA and 180-bp repeats. The sequence continues to be extended further into centromeric and other regions of complex sequence.

Characterization of the coding regions

To assess the similarities and differences of the *Arabidopsis* gene complement compared with other sequenced eukaryotic genomes, we assigned functional categories to the complete set of *Arabidopsis* genes. For chromosome 4 genes and the yeast genome, predicted functions were previously manually assigned[5,21]. All other predicted proteins were automatically assigned to these functional categories[22], assuming that conserved sequences reflect common functional relationships.

The functions of 69% of the genes were classified according to sequence similarity to proteins of known function in all organisms; only 9% of the genes have been characterized experimentally (Fig. 2a). Generally similar proportions of gene products were predicted to be targeted to the secretory pathway and mitochondria in *Arabidopsis* and yeast, and up to 14% of the gene products are

Table 1 Summary statistics of the *Arabidopsis* genome

Feature	Value					
(a) The DNA molecules	Chr. 1	Chr. 2	Chr. 3	Chr. 4	Chr. 5	Σ
Length (bp)	29,105,111	19,646,945	23,172,617	17,549,867	25,953,409	115,409,949
Top arm (bp)	14,449,213	3,607,091	13,590,268	3,052,108	11,132,192	
Bottom arm (bp)	14,655,898	16,039,854	9,582,349	14,497,759	14,803,217	
Base composition (%GC)						
Overall	33.4	35.5	35.4	35.5	34.6	
Coding	44.0	44.0	44.3	44.1	44.1	
Non-coding	32.4	32.9	33.0	32.8	32.5	
Number of genes	6,543	4,036	5,220	3,825	5,874	25,498
Gene density (kb per gene)	4.0	4.9	4.6	4.6	4.4	
Average gene length (bp)	2,078	1,949	1,925	2,136	1,974	
Average peptide length (bp)	446	421	424	448	429	
Exons						
Number	35,482	19,631	26,570	20,073	31,226	132,982
Total length (bp)	8,772,569	5,100,288	6,654,507	5,150,883	7,571,013	33,249,250
Average per gene	5.4	4.9	5.1	5.2	5.3	
Average size (bp)	247	259	250	256	242	
Introns						
Number	28,939	15,595	21,350	16,248	25,352	107,484
Total length (bp)	4,828,768	2,788,430	3,397,531	3,030,649	4,030,045	18,055,421
Average size (bp)	168	177	159	186	159	
Number of genes with ESTs (%)	60.8	56.9	59.8	61.4	61.4	
Number of ESTs	30,522	14,989	20,782	16,605	22,895	105,793
(b) The proteome						
Classification/function						
Total proteins	6,543	4,036	5,220	3,825	5,874	25,498
With INTERPRO	4,194	1,205	2,989	1,545	3,136	13,069
domains	64.1%	29.9%	57.8%	40.4%	53.4%	51.3%
Genes containing at	2,334	1,322	1,615	1,402	1,940	8,613
least one TM domain	35.7%	32.8%	30.9%	36.7%	33.0%	33.8%
Genes containing at	2,513	1,424	1,664	1,304	2,121	9,026
least one SCOP domain	38.4%	35.3%	31.9%	34.1%	36.1%	35.4%
With putative signal peptides						
Secretory pathway	1,242 19.0%	675 16.7%	877 17.0%	659 17.2%	1,014 17.3%	4,467 17.6%
>0.95 specificity	1,146 17.5%	632 15.7%	813 15.7%	632 16.5%	964 16.4%	4,187 16.4%
Chloroplast	866 13.2%	535 13.2%	754 14.6%	532 13.9%	887 15.1%	3,574 14.0%
>0.95 specificity	602 9.2%	290 7.2%	420 8.1%	298 7.8%	475 8.1%	2,085 8.2%
mitochondria	901 13.8%	425 10.5%	554 10.7%	390 10.2%	627 10.7%	2,897 11.4%
>0.95 specificity	113 1.7%	49 1.2%	63 1.2%	59 1.5%	65 1.1%	349 1.4%
Functional classification						
Cellular metabolism	1,188 22.7%	620 23.3%	745 22.8%	588 22.9%	868 21.1%	4,009 22.5%
Transcription	880 16.8%	474 17.8%	566 17.3%	335 13.1%	763 18.6%	3,018 16.9%
Plant defence	640 12.2%	276 10.4%	354 10.8%	295 11.5%	490 11.9%	2,055 11.5%
Signalling	573 11.0%	296 11.1%	356 10.9%	210 8.2%	420 10.2%	1,855 10.4%
Growth	542 10.4%	263 9.9%	357 10.9%	448 17.5%	469 11.4%	2,079 11.7%
Protein fate	520 9.9%	273 10.2%	314 9.6%	264 10.3%	395 9.6%	1,766 9.9%
Intracellular transport	435 8.3%	214 8.9%	269 8.2%	220 8.6%	334 8.1%	1,472 8.3%
Transport	236 4.5%	139 5.2%	155 4.7%	113 4.4%	206 5.0%	849 4.8%
Protein synthesis	216 4.1%	111 4.2%	148 4.5%	90 3.5%	165 4.0%	730 4.1%
Total	5,230	2,666	3,264	2,563	4,110	17,833

The features of *Arabidopsis* chromosomes 1–5 and the complete nuclear genome are listed. Specialized searches used the following programs and databases: INTERPRO[23]; transmembrane (TM) domains by ALOM2 (unpublished); SCOP domain database[121]; functional classification by the PEDANT analysis system[22]. Signal peptide prediction (secretory pathway, targeted to chloroplast or mitochondria) was performed using TargetP[122] and http://www.cbs.dtu.dk/services/TargetP/.
*Default value.

likely to be targeted to the chloroplast (Table 1). The significant proportion of genes with predicted functions involved in metabolism, gene regulation and defence is consistent with previous analyses[5]. Roughly 30% of the 25,498 predicted gene products, (Fig. 2a), comprising both plant-specific proteins and proteins with similarity to genes of unknown function from other organisms, could not be assigned to functional categories.

To compare the functional catagories in more detail, we compared data from the complete genomes of *Escherichia coli*[23], *Synechocystis* sp.[24], *Saccharomyces cerevisiae*[21], *C. elegans*[1] and *Drosophila*[2], and a non-redundant protein set of *Homo sapiens*, with the *Arabidopsis* genome data (Fig. 2b), using a stringent BLASTP threshold value of $E < 10^{-30}$. The proportion of *Arabidopsis* proteins having related counterparts in eukaryotic genomes varies by a factor of 2 to 3 depending on the functional category. Only 8–23% of *Arabidopsis* proteins involved in transcription have related genes in other eukaryotic genomes, reflecting the independent evolution of many plant transcription factors. In contrast, 48–60% of genes involved in protein synthesis have counterparts in the other eukaryotic genomes, reflecting highly conserved gene functions. The relatively high proportion of matches between *Arabidopsis* and bacterial proteins in the categories 'metabolism' and 'energy' reflects both the acquisition of bacterial genes from the ancestor of the plastid and high conservation of sequences across all species. Finally, a comparison between unicellular and multicellular eukaryotes indicates that *Arabidopsis* genes involved in cellular communication and signal transduction have more counterparts in multicellular eukaryotes than in yeast, reflecting the need for sets of genes for communication in multicellular organisms.

Pronounced redundancy in the *Arabidopsis* genome is evident in segmental duplications and tandem arrays, and many other genes with high levels of sequence conservation are also scattered over the genome. Sequence similarity exceeding a BLASTP value $E < 10^{-20}$ and extending over at least 80% of the protein length were used as parameters to identify protein families (Table 2). A total of 11,601 protein types were identified. Thirty-five per cent of the predicted proteins are unique in the genome, and the proportion of proteins belonging to families of more than five members is substantially higher in *Arabidopsis* (37.4%) than in *Drosophila* (12.1%) or

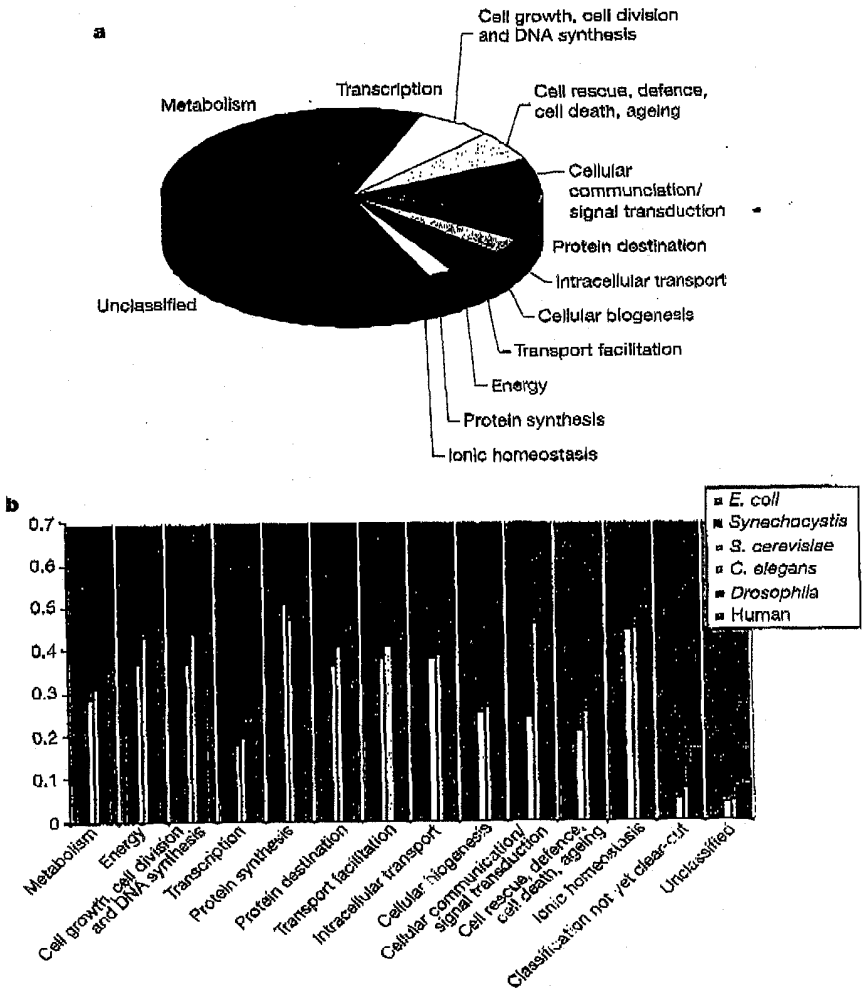

Figure 2 Functional analysis of *Arabidopsis* genes. **a**, Proportion of predicted *Arabidopsis* genes in different functional categories. **b**, Comparison of functional categories between organisms. Subsets of the *Arabidopsis* proteome containing all proteins that fall into a common functional class were assembled. Each subset was searched against the complete set of translations from *Escherichia coli*, *Synechocystis* sp. PCC6803, *Saccharomyces cerevisae*, *Drosophila*, *C. elegans* and a *Homo sapiens* non-redundant protein database. The percentage of *Arabidopsis* proteins in a particular subset that had a BLASTP match with $E \leq 10^{-30}$ to the respective reference genome is shown. The y axis reflects the measure of sequence conservation of proteins within this particular functional category between *Arabidopsis* and the respective reference genome. y axis, 0.1 = 10%.

Table 2 Proportion of genes in different organisms present as either singletons or in paralogous families

	No of singletons and distinct gene families	Unique	Gene families containing				
			2 members	3 members	4 members	5 members	>5 members
H. influenzae	1,587	88.8%	6.8%	2.3%	0.7%	0.0%	1.4%
S. cerevisiae	5,105	71.4%	13.8%	3.5%	2.2%	0.7%	8.4%
D. melanogaster	10,736	72.5%	8.5%	3.4%	1.9%	1.6%	12.1%
C. elegans	14,177	55.2%	12.0%	4.5%	2.7%	1.6%	24.0%
Arabidopsis	11,601	35.0%	12.5%	7.0%	4.4%	3.6%	37.4%

The number of genes in the genomes of *Haemophilus influenzae*, *S. cerevisiae*, *Drosophila*, *C. elegans* and *Arabidopsis* that are present either as singletons or in gene families with two or more members are listed. To be grouped in a gene family, two genes had to show similarity exceeding a BLASTP value $E < 10^{-20}$ and a FASTA alignment over at least 80% of the protein length. In column 1, the number of genes that are unique plus the number of gene families are listed. Columns 2 to 6 give the percentage of genes present as singletons or in gene families of *n* members.

C. elegans (24.0%). The absolute number of *Arabidopsis* gene families and singletons (types) is in the same range as the other multicellular eukaryotes, indicating that a proteome of 11,000–15,000 types is sufficient for a wide diversity of multicellular life. The proportion of gene families with more than two members is considerably more pronounced in *Arabidopsis* than in other eukaryotes (Fig. 3). As segmental duplication is responsible for 6,303 gene duplications (see below), the extent of tandem gene duplications accounts for a significant proportion of the increased family size. These features of the *Arabidopsis*, and presumably other plant genomes, may indicate more relaxed constraints on genome size in plants, or a more prominent role of unequal crossing over to generate new gene copies.

Conserved protein domains revealed more informative differences through INTERPRO[25] analysis of the predicted gene products from *Arabidopsis*, *S. cerevisiae*, *C. elegans* and *Drosophila*. Statistically over-represented domains, and those that are absent from the *Arabidopsis* genome, indicate domains that may have been gained or lost during the evolution of plants (Supplementary Information Table 1). Proteins containing the Pro-Pro-Arg repeat, which is involved in RNA stabilization and RNA processing, are over-represented as compared to yeast, fly and worm; 400 proteins containing this signature were detected in *Arabidopsis* compared with only 10 in total in yeast, *Drosophila* and *C. elegans*. Protein kinases and associated domains, 169 proteins containing a disease resistance protein signature, and the Toll/IL-1R (TIR) domain, a component of pathogen recognition molecules[26], are also relatively abundant. This suggests that pathways transducing signals in response to pathogens and diverse environmental cues are more abundant in plants than in other organisms.

The RING zinc finger domain is relatively over-represented in *Arabidopsis* compared with yeast, *Drosophila* and *C. elegans*, whereas the F-box domain is over-represented as compared with yeast and *Drosophila* only. These domains are involved in targeting proteins to the proteasome[27] and ubiquitinylation[28] pathways of protein degradation, respectively. In plants many processes such as hormone and defence responses, light signalling, and circadian rhythms and pattern formation use F-box function to direct negative regulators to the ubiquitin degradation pathway. This mode of regulation appears to be more prevalent in plants and may account for a higher representation of the F box than in *Drosophila* and for the over-representation of the ubiquitin domain in the *Arabidopsis* genome. RING finger domain proteins in general have a role in ubiquitin protein ligases, indicating that proteasome-mediated degradation is a more widespread mode of regulation in plants than in other kingdoms.

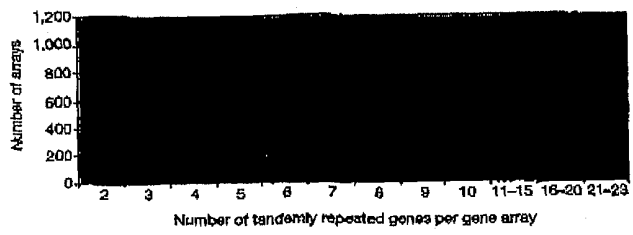

Figure 3 Distribution of tandemly repeated gene arrays in the *Arabidopsis* genome. Tandemly repeated gene arrays were identified using the BLASTP program with a threshold of $E < 10^{-20}$. One unrelated gene among cluster members was tolerated. The histogram gives the number of clusters in the genome containing 2 to *n* similar gene units in tandem.

Most functions identified by protein domains are conserved in similar proportions in the *Arabidopsis*, *S. cerevisiae*, *Drosophila* and *C. elegans* genomes, pointing to many ubiquitous eukaryotic pathways. These are illustrated by comparing the list of human disease genes[29] to the complete *Arabidopsis* gene set using BLASTP. Out of 289 human disease genes, 139 (48%) had hits in *Arabidopsis* using a BLASTP threshold $E < 10^{-10}$. Sixty-nine (24%) exceeded an $E < 10^{-40}$ threshold, and 26 (9.3%) had scores better than $E < 10^{-100}$ (Table 3). There are at least 17 human disease genes more similar to *Arabidopsis* genes than yeast, *Drosophila* or *C. elegans* genes (Table 3).

This analysis shows that, although numerous families of proteins are shared between all eukaryotes, plants contain roughly 150 unique protein families. These include transcription factors, structural proteins, enzymes and proteins of unknown function. Members of the families of genes common to all eukaryotes have undergone substantial increases or decreases in their size in *Arabidopsis*. Finally, the transfer of a relatively small number of cyanobacteria-related genes from a putative endosymbiotic ancestor of the plastid has added to the diversity of protein structures found in plants.

Genome organization and duplication

The *Arabidopsis* genome sequence provides a complete view of chromosomal organization and clues to its evolutionary history. Gene families organized in tandem arrays of two or more units have been described in *C. elegans*[1] and *Drosophila*[2]. Analysis of the *Arabidopsis* genome revealed 1,528 tandem arrays containing 4,140 individual genes, with arrays ranging up to 23 adjacent members (Fig. 3). Thus 17% of all genes of *Arabidopsis* are arranged in tandem arrays.

Large segmental duplications were identified either by directly aligning chromosomal sequences or by aligning proteins and searching for tracts of conserved gene order. All five chromosomes were aligned to each other in both orientations using MUMmer[30], and the results were filtered to identify all segments at least 1,000 bp in length with at least 50% identity (Supplementary Information Fig. 1). These revealed 24 large duplicated segments of 100 kb or larger, comprising 65.6 Mb or 58% of the genome. The only duplicated segment in the centromeric regions was a 375-kb segment on chromosome 4. Many duplications appear to have undergone further shuffling, such as local inversions after the duplication event.

We used TBLASTX[3] to identify collinear clusters of genes residing in large duplicated chromosomal segments. The duplicated regions encompass 67.9 Mb, 60% of the genome, slightly more than was

found in the DNA-based alignment (Fig. 4), and these data extend earlier findings[4,5,31]. The extent of sequence conservation of the duplicated genes varies greatly, with 6,303 (37%) of the 17,193 genes in the segments classified as highly conserved ($E < 10^{-30}$) and a further 1,705 (10%) showing less significant similarity up to $E < 10^{-5}$. The proportion of homologous genes in each duplicated segment also varies widely, between 20% and 47% for the highly conserved class of genes. In many cases, the number of copies of a gene and its counterpart differ (for example, one copy on one chromosome and multiple copies on the other; see Supplementary Information Fig. 2); this could be due to either tandem duplication or gene loss after the segmental duplication.

What does the duplication in the *Arabidopsis* genome tell us about the ancestry of the species? Polyploidy occurs widely in plants and is proposed to be a key factor in plant evolution[32]. As the majority of the *Arabidopsis* genome is represented in duplicated (but not triplicated) segments, it appears most likely that *Arabidopsis*, like maize, had a tetraploid ancestor[33]. A comparative sequence analysis of *Arabidopsis* and tomato estimated that a duplication occurred ~112 Myr ago to form a tetraploid[34]. The degrees of conservation of the duplicated segments might be due to divergence from an ancestral autotetraploid form, or might reflect differences present in an allotetraploid ancestor. It is also possible, however, that several independent segmental duplication events took place instead of tetraploid formation and stabilization.

The diploid genetics of *Arabidopsis* and the extensive divergence of the duplicated segments have masked its evolutionary history. The determination of *Arabidopsis* gene functions must therefore be pursued with the potential for functional redundancy taken into account. The long period of time over which genome stabilization has occurred has, however, provided ample opportunity for the divergence of the functions of genes that arose from duplications.

Comparative analysis of *Arabidopsis* accessions

Comparing the multiple accessions of *Arabidopsis* allows us to identify commonly occurring changes in genome microstructure. It also enables the development of new molecular markers for genetic mapping. High rates of polymorphism between *Arabidopsis* accessions, including both DNA sequence and copy number of tandem arrays, are prevalent at loci involved in disease resistance[35]. This has been observed for other plant species, and such loci are thought to serve as templates for illegitimate recombination to create new pathogen response specificities[36]. We carried out a comparative analysis between 82 Mb of the genome sequence of *Arabidopsis* accession Columbia (Col-0) and 92.1 Mb of non-redundant low-pass (twofold redundant) sequence data of the genomic DNA of accession Landsberg *erecta* (*Ler*). We identified two classes of differences between the sequences: single nucleotide polymorphisms (SNPs), and insertion–deletions (InDels). As we used high stringency criteria, our results represent a minimum estimate of numbers of polymorphisms between the two genomes.

In total, we detected 25,274 SNPs, representing an average density of 1 SNP per 3.3 kb. Transitions (A/T–G/C) represented 52.1% of the SNPs, and transversions accounted for the remainder: 17.3% for A/T–T/A, 22.7% for A/T–C/G and 7.9% for C/G–G/C. In total, we detected 14,570 InDels at an average spacing of 6.1 kb. They ranged from 2 bp to over 38 kilobase-pairs, although 95% were smaller than 50 bp. Only 10% of the InDels were co-located with simple sequence repeats identified with the program Sputnik. An analysis of 416 relative insertions greater than 250 bp in Col-0 showed that 30% matched transposon-related proteins, indicating that a substantial proportion of the large InDels are the result of transposon insertion or excision. Many InDels contained entire active genes not related to transposons. Half of such genes absent from corresponding positions in the Col-0 sequence were found elsewhere on the genome of Ler. This indicates that genes have been transferred to new genomic locations.

Gene structures are often affected by small InDels and SNPs. The positions of SNPs and InDels were mapped relative to 87,427 exons and 70,379 introns annotated in the Col-0 sequence. SNPs were found in exons, introns and intergenic regions at frequencies of 1 SNP per 3.1, 2.2 and 3.5 kb, respectively. The frequencies for InDels were 1 per 9.3, 3.1 and 4.3 kb, respectively. Polymorphisms were detected in 7% of exons, and alter the spliced sequences of 25% of the predicted genes. For InDels in exons, insertion lengths divisible by three are prevalent for small insertions (< 50 bp), indicating that many proteins can withstand small insertions or deletions of amino acids without loss of function.

Our analyses show that sequence polymorphisms between accessions of *Arabidopsis* are common, and that they occur in both coding and non-coding regions. We found evidence for the relocation of genes in the genome, and for changes in the complement of transposable elements. The data presented here are available at http://www.arabidopsis.org/cereon/.

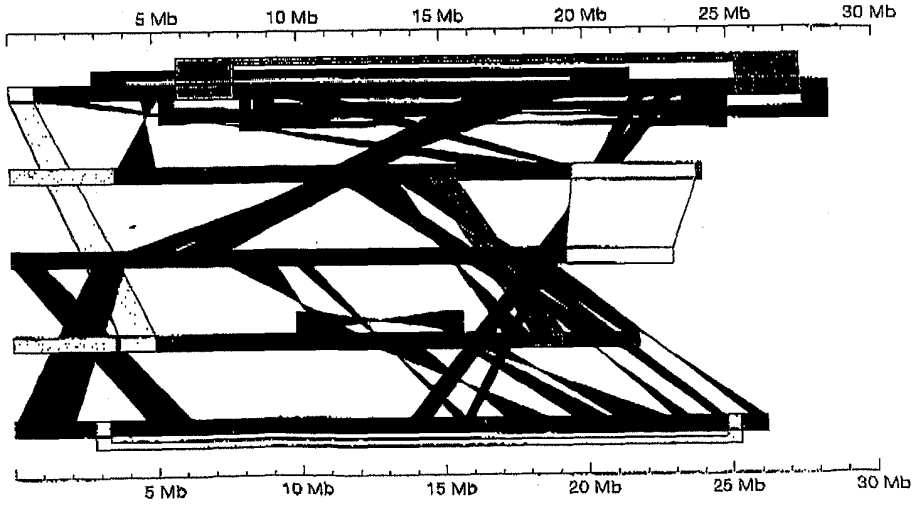

Figure 4 Segmentally duplicated regions in the *Arabidopsis* genome. Individual chromosomes are depicted as horizontal grey bars (with chromosome 1 at the top), centromeres are marked black. Coloured bands connect corresponding duplicated segments. Similarity between the rDNA repeats are excluded. Duplicated segments in reversed orientation are connected with twisted coloured bands. The scale is in megabases.

Comparison of *Arabidopsis* and other plant genera

Comparative genetic mapping can reveal extensive conservation of genome organization between closely related species[37,38]. The comparative analysis of plant genome microstructure reveals much about the evolution of plant genomes and provides unprecedented opportunities for crop improvement by establishing the detailed structures of, and relationships between, the genomes of crops and *Arabidopsis*.

The lineages leading to *Arabidopsis* and *Capsella rubella* (shepherd's purse) diverged between 6.2 and 9.8 Myr ago, and the gene content and genome organization of *C. rubella* is very similar to that of *Arabidopsis*[39], including the large-scale duplications. Alignment of *Arabidopsis* complementary DNA and EST sequences with genomic DNA sequences of *Arabidopsis* and *C. rubella* showed conservation of exon length and intron positions. Coding sequences predicted from these alignments differed from the annotated *Arabidopsis* gene sequences in two out of five cases.

The ancestral lineages of *Arabidopsis* and the *Brassica* (cabbage and mustard) genera diverged 12.2–19.2 Myr ago[40]. *Brassica* genes show a high level of nucleotide conservation with their *Arabidopsis* orthologues, typically more than 85% in coding regions[40]. The structure of *Brassica* genomes resembles that of *Arabidopsis*, but with extensive triplication and rearrangement[41], and extensive divergence of microstructure (Supplementary Information Fig. 3). The divergence between the genomes of *Arabidopsis* and *Brassica oleracea* is in striking contrast to that observed between *Arabidopsis* and *C. rubella*, although the time since divergence is only twofold greater. This accelerated rate of change in triplicated segments of the genome of *B. oleracea* indicates that polyploidy fosters rapid chromosomal evolution.

The *Arabidopsis* and tomato lineages diverged roughly 150 Myr ago, and comparative sequence analysis of segments of their genomes has revealed complex relationships[34]. Four regions of the *Arabidopsis* genome are related to each other and to one region in the tomato genome, suggesting that two rounds of duplication may have occurred in the *Arabidopsis* lineage. The extensive duplication described here supports the proposal that the more recent of these duplications, estimated to have occurred ~112 Myr ago, was the result of a polyploidization event. The lineages of *Arabidopsis* and rice diverged ~200 Myr ago[42]. Three regions of the genome of *Arabidopsis* were related to each other and to one region in the rice genome, providing further evidence for multiple duplication events[43,44].

The frequent occurrence of tandem gene duplications and the apparent deletion of single genes, or small groups of adjacent genes, from duplicated regions suggests that unequal crossing over may be a key mechanism affecting the evolution of plant genome microstructure. However, the segmental inversions and gene translocations in the genomes of both rice and *B. oleracea* that are not found in *Arabidopsis* indicate that additional mechanisms may be involved[40].

Integration of the three genomes in the plant cell

The three genomes in the plant cell—those of the nucleus, the plastids (chloroplasts) and the mitochondria—differ markedly in gene number, organization and stability. Plastid genes are densely packed in an order highly conserved in all plants[45], whereas mitochondrial genes[46] are widely dispersed and subjected to extensive recombination.

Organellar genomes are remnants of independent organisms—plastids are derived from the cyanobacterial lineage and mitochondria from the α-Proteobacteria. The remaining genes in plastids include those that encode subunits of the photosystem and the electron transport chain, whereas the genes in mitochondria encode essential subunits of the respiratory chain. Both organelles contain sets of specific membrane proteins that, together with housekeeping proteins, account for 61% of the genes in the chloroplast and 88 % in the mitochondrion (Table 4). The balances are involved in transcription and translation.

The number of proteins encoded in the nucleus likely to be found

Table 3 *Arabidopsis* genes with similarities to human disease genes

Human disease gene	*E* value	Gene code	*Arabidopsis* hit
Darier–White, SERCA	5.9×10^{-272}	T27I1_16	Putative calcium ATPase
Xeroderma Pigmentosum, D-XPD	7.2×10^{-228}	F15K9_19	Putative DNA repair protein
Xeroderma pigment, B-ERCC3	9.6×10^{-214}	AT5g41360	DNA excision repair cross-complementing protein
Hyperinsulinism, ABCC8	7.1×10^{-165}	F20D22_11	Multidrug resistance protein
Renal tubul. acidosis, ATP6B1	1.0×10^{-162}	AT4g38510	Probable H+-transporting ATPase
HDL deficiency 1, ABCA1	2.4×10^{-161}	At2g41700	Putative ABC transporter
Wilson, ATP7B	7.6×10^{-161}	AT5g44790	ATP-dependent copper transporter
Immunodeficiency, DNA Ligase 1	8.2×10^{-172}	T6D22_10	DNA ligase
Stargardt's, ABCA4	2.8×10^{-168}	At2g41700	Putative ABC transporter
Ataxia telangiectasia, ATM	3.1×10^{-158}	AT3g48190	Ataxia telangiectasia mutated protein AtATM
Niemann–Pick, NPC1	1.2×10^{-166}	F7F22_1	Niemann–Pick C disease protein-like protein
Menkes, ATP7A	1.1×10^{-153}	F2K11_17	ATP-dependent copper transporter, putative
HNPCC*, MLH1	1.5×10^{-150}	AT4g09140	MLH1 protein
Deafness, hereditary, MYO15	2.7×10^{-150}	At2g31900	Putative unconventional myosin
Fam. cardiac myopathy, MYH7	6.5×10^{-147}	T1G11_14	Putative myosin heavy chain
Xeroderma Pigmentosum, F-XPF	1.4×10^{-146}	AT5g41150	Repair endonuclease (gb|AAF01274.1)
G6PD deficiency, G6PD	7.6×10^{-137}	AT5g40760	Glucose-6-phosphate dehydrogenase
Cystic fibrosis, ABCC7	2.3×10^{-135}	AT3g62700	ABC transporter-like protein
Glycerol kinase defic, GK	7.9×10^{-135}	T21F11_21	Putative glycerol kinase
HNPCC, MSH3	6.6×10^{-134}	AT4g25540	Putative DNA mismatch repair protein
HNPCC, PMS2	5.1×10^{-128}	AT4g02460	No title
Zellweger, PEX1	4.1×10^{-126}	AT5g08470	Putative protein
HNPCC, MSH6	9.6×10^{-122}	AT4g02070	G/T DNA mismatch repair enzyme
Bloom, BLM	4.4×10^{-109}	T19D16_15	DNA helicase isolog
Finnish amyloidosis, GSN	2.2×10^{-107}	AT5g57320	Villin
Chediak–Higashi, CHS1	5.8×10^{-99}	F10O3_11	Putative transport protein
Xeroderma Pigmentosum, G-XPG	7.1×10^{-99}	AT3g28030	Hypothetical protein
Bare lymphocyte, ABCB3	1.3×10^{-84}	AT5g39040	ABC transporter-like protein
Citrullinemia, type I, ASS	3.2×10^{-83}	AT4g24830	Argininosuccinate synthase-like protein
Coffin–Lowry, RPS6KA3	5.2×10^{-81}	AT3g08720	Putative ribosomal-protein S6 kinase (ATPK19)
Keratoderma, KRT9	8.5×10^{-81}	AT3g17050	Unknown protein
Myotonic dystrophy, DM1	1.4×10^{-76}	At2g20470	Putative protein kinase
Bartter's, SLC12A1	1.6×10^{-75}	F26G16_9	Cation-chloride co-transporter, putative
Dents, CLCN5	3.3×10^{-74}	AT5g26240	CLC-d chloride channel protein
Diaphanous 1, DAPH1	1.9×10^{-73}	68069_m00158	Hypothetical protein
AKT2	6.9×10^{-72}	AT3g08730	Putative ribosomal-protein S6 kinase (ATPK6)

in organelles was predicted using default settings on TargetP (Table 1). Many nuclear gene products that are targeted to either (or both) organelles were originally encoded in the organelle genomes and were transferred to the nuclear genome during evolutionary history. A large number also appear to be of eukaryotic origin, with functions such as protein import components, which were probably not required by the free-living ancestors of the endosymbionts.

To identify nuclear genes of possible organellar ancestry, we compared all predicted *Arabidopsis* proteins to all proteins from completed genomes including those from plastids and mitochondria (Supplementary Information Table 2). This search identified proteins encoded by the *Arabidopsis* nuclear genome that are most similar to proteins encoded by other species' organelle genomes (14 mitochondrial and 44 plastid). These represent organelle-to-nuclear gene transfers that have occurred sometime after the divergence of the organelle-containing lineages[47]. There is a great excess of nuclear encoded proteins most similar to proteins from the cyanobacteria *Synechocystis* (Supplementary Information Fig. 4; 806 *Arabidopsis* predicted proteins matching 404 different *Synechocystis* proteins, providing further evidence of a genome duplication). These 806 *Arabidopsis* predicted proteins, and many others of greatly diverse function, are possibly of plastid descent. Through searches against proteins from other cyanobacteria (with incompletely sequenced genomes), we identified 69 additional genes of possibly plastid descent. Only 25% of these putatively plastid-derived proteins displayed a target peptide predicted by TargetP, indicating potential cytoplasmic functions for most of these genes.

The difference between predicted plastid-targeted and predicted plastid-derived genes indicates that there is a probable overestimation by *ab initio* targeting prediction methods and a lack of resolution with respect to destination organelles, the possible extensive divergence of some endosymbiont-derived genes in the nuclear genome, the co-opting of nuclear genes for targeting to organelles, and cytoplasmic functions for cyanobacteria-derived proteins. Clearly more refined tools and extensive experimentation is required to catalogue plastid proteins.

The transfer of genes between genomes still continues (Supplementary Information Table 3). Plastid DNA insertions in the nucleus (17 insertions totalling 11 kb) contain full-length genes encoding proteins or tRNAs, fragments of genes and an intron as well as intergenic regions. Subsequent reshuffling in the nucleus is illustrated by the *atpH* gene, which was originally transferred completely, but is now in two pieces separated by 2 kb. The 13 small mitochondrial DNA insertions total 7 kb in addition to the large insertion close to the centromere of chromosome 2 (ref. 3). The high level of recombination in the mitochondrial genome may account for these events.

Transposable elements

Transposons, which were originally identified in maize by Barbara McClintock, have been found in all eukaryotes and prokaryotes. A subset of transposons replicate through an RNA intermediate (class I), whereas others move directly through a DNA form (class II). Transposons are further classified by similarity either between their mobility genes or between their terminal and/or internal motifs, as well as by the size and sequence of their target site. Internally deleted elements can often be mobilized in *trans* by fully functional elements.

Transposons in *Arabidopsis* account for at least 10% of the genome, or about one-fifth of the intergenic DNA. The *Arabidopsis* genome has a wealth of class I (2,109) and II (2,203) elements, including several new groups (1,209 elements; Supplementary Information Table 4). Mobile histories for many elements were obtained by identifying regions of the genome with significant similarity to 'empty' target sites (RESites) thus providing high-resolution information concerning the termini and target site duplications[48,49]. These regions were readily detected because of the propensity of transposons to integrate into repeats and because of duplications in the genome sequence. In several cases, genes appear to have been included as 'passengers' in transposable units[48]. In some cases, shared sequence similarity, coding capacity and RESites attest to recent activity of transposable elements in the *Arabidopsis* genome. Only about 4% of the complete elements identified correspond to an EST, however, suggesting that most are not transcribed.

Transposable elements found in many other plant genomes are well represented in *Arabidopsis*, including *copia*- and *gypsy*-like long terminal repeat (LTR) retrotransposons, long interspersal nuclear elements (LINEs); short interspersed nuclear elements (SINEs), *hobo*/*Activator*/Tam3 (*hAT*)-like elements, CACTA-like elements and miniature inverted-repeat transposable elements (MITES). Although usually small in size, some larger *Tourist*-like MITEs contain open reading frames (ORFs) with similarity to the transposases of bacterial insertion sequences[48]. *Basho* and many *Mutator*-like elements (MULEs), first discovered in the *Arabidopsis* sequence, represent structurally unique transposons[48–50]. *Basho* elements have a target site preference for mononucleotide 'A' and wide distribution among plants[48,51]. MULEs exhibit a high level of sequence diversity and members of most groups lack long terminal inverted repeats (TIRs). Phylogenetic analysis of the *Arabidopsis* MURA-like transposases suggests that TIR-containing MULEs are more closely related to one another than to MULEs lacking TIRs[49,52].

For many plants with large genomes, class I retrotransposons contribute most of the nucleotide content[53]. In the small *Arabidopsis* genome, class I elements are less abundant and primarily occupy the centromere. In contrast, *Basho* elements and class II transposons such as MITEs and MULEs predominate on the periphery of pericentromeric domains (Fig. 5). In class II transposons, MULEs and CACTA elements are clustered near centromeres and heterochromatic knobs, whereas MITEs and *hAT* elements have a less pronounced bias. The distribution pattern of transposable elements observed in *Arabidopsis* may reflect different types of pericentromeric heterochromatin regions and may be similar to those found in animals.

Numerous centromeric satellite repeats are located between each chromosome arm and have not yet been sequenced, but are represented in part by unanchored BAC contigs (R. Martienssen and M. Marra, unpublished data). End sequence suggests that these domains contain many more class I than class II elements, consistent with the distribution reported here (K. Lemcke and R. Martienssen, unpublished data). We do not know the significance of the apparent paucity of elements in telomeric regions and in the region flanking the rDNA repeats on chromosome 4 (but not on chromosome 2).

Overall, transposon-rich regions are relatively gene-poor and have lower rates of recombination and EST matches, indicating a correlation between low gene expression, high transposon density and low recombination[51]. The role of transposons in genome

Table 4 General features of genes encoded by the three genomes in *Arabidopsis*

	Nucleus/cytoplasm	Plastid	Mitochondria
Genome size	125 Mb	154 kb	367 kb
Genome equivalent/cell	2	560	26
Duplication	60%	17%	10%
Number of protein genes	25,498	79	58
Gene order	Variable, but syntenic	Conserved	Variable
Density (kb per protein gene)	4.5	1.2	6.25
Average coding length	1,900 nt	900 nt	860 nt
Genes with introns	79%	18.4%	12%
Genes/pseudogenes	1/0.03	1/0	1/0.2–0.5
Transposons (% of total genome size)	14%	0%	4%

organization and chromosome structure can now be addressed in a model organism known to undergo DNA methylation and other forms of chromatin modification thought to regulate transposition[52].

rDNA, telomeres and centromeres

Nucleolar organizers (NORs) contain arrays of unit repeats encoding the 18S, 5.8S and 25S ribosomal RNA genes and are transcribed by RNA polymerase I. Together with 5S RNA, which is transcribed by RNA polymerase III, these rRNAs form the structural and catalytic cores of cytoplasmic ribosomes. In *Arabidopsis*, the NORs juxtapose the telomeres of chromosomes 2 and 4, and comprise uninterrupted 18S, 5.8S and 25S units all orientated on the chromosomes in the same direction[54]. In contrast, the 5S rRNA genes are localized to heterogeneous arrays in the centromeric regions of chromosomes 3, 4 and 5 (ref. 55; and Fig. 6). Both NORs are roughly 3.5–4.0 megabase-pairs and comprise ~350–400 highly methylated rRNA gene units, each ~10 kb (ref. 54). The sequence between the euchromatic arms and NORs has been determined. Elsewhere in the genome, only one other 18S, 5.8S, 25S rRNA gene unit was identified in centromere 3. Although minor variations in sequence length and composition occur in the NOR repeats, these variants are highly clustered, supporting a model of sequence maintenance through concerted evolution[55].

Arabidopsis telomeres are composed of CCCTAAA repeats and average ~2–3 kb (ref. 56). For TEL4N (telomere 4 North), consensus repeats are adjacent to the NOR; the remaining telomeres are typically separated from coding sequences by repetitive subtelomeric regions measuring less than 4 kb. Imperfect telomere-like arrays of up to 24 kb are found elsewhere in the genome, particularly near centromeres. These arrays might affect the expression of nearby genes and may have resulted from ancient rearrangements, such as inversions of the chromosome arms.

Centromere DNA mediates chromosome attachment to the meiotic and mitotic spindles and often forms dense heterochromatin. Genetic mapping of the regions that confer centromere function provided the markers necessary to precisely place BAC clones at individual centromeres[17]; 69 clones were targeted for sequencing, resulting in over 5 Mb of DNA sequence from the centromeric regions. The unsequenced regions of centromeres are composed primarily of long, homogeneous arrays that were characterized previously with physical[57] and genetic mapping[17] and contain over 3 Mb of repetitive arrays, including the 180-bp repeats and 5S rDNA[51] (Fig. 6).

Arabidopsis centromeres, like those of many higher eukaryotes, contain numerous repetitive elements including retroelements, transposons, microsatellites and middle repetitive DNA[17]. These repeats are rare in the euchromatic arms and often most abundant in pericentromeric DNA. The repeats, affinity for DNA-binding dyes, dense methylation patterns and inhibition of homologous recombination indicate that the centromeric regions are highly heterochromatic, and such regions are generally viewed as very poor environments for gene expression. Unexpectedly, we found at least 47 expressed genes encoded in the genetically defined centromeres of *Arabidopsis* (http://preuss.bsd.uchicago.edu/arabidopsis.genome.html). In several cases, these genes reside on islands of unique sequence flanked by repetitive arrays, such as 180-bp or 5S rDNA repeats. Among the genes encoded in the centromeres are members of 11 of the 16 functional categories that comprise the proteome. The centromeres are not subject to recombination; consequently, genes residing in these regions probably exhibit unique patterns of molecular evolution.

The function of higher eukaryotic centromeres may be specified by proteins that bind to centromere DNA, by epigenetic modifications, or by secondary or higher order structures. A pairwise comparison of the non-repetitive portions of all five centromeres showed they share limited (1–7%) sequence similarity. Forty-one families of small, conserved centromere sequences (AtCCS, see http://preuss.bsd.uchicago.edu/arabidopsis.genome.html) are enriched in the centromeric and pericentromeric regions and differ from sequences found in the centromeres of other eukaryotes. Molecular and genetic assays will be required to determine whether these conserved motifs nucleate *Arabidopsis* centromere activity. Apart from the AtCCS sequences, most centromere DNA is not shared between chromosomes, complicating efforts to derive clear evolutionary relationships. In contrast, genetic and cytological assays indicate that homologous centromeres are highly conserved among *Arabidopsis* accessions, albeit subject to rearrangements such as inversions to form knobs[5,58,59] and insertions[4]. Further investigation of centromere DNA promises to yield information on the evolutionary forces that act in regions of limited recombination, as well as an improved understanding of the role of DNA sequence patterns in chromosome segregation.

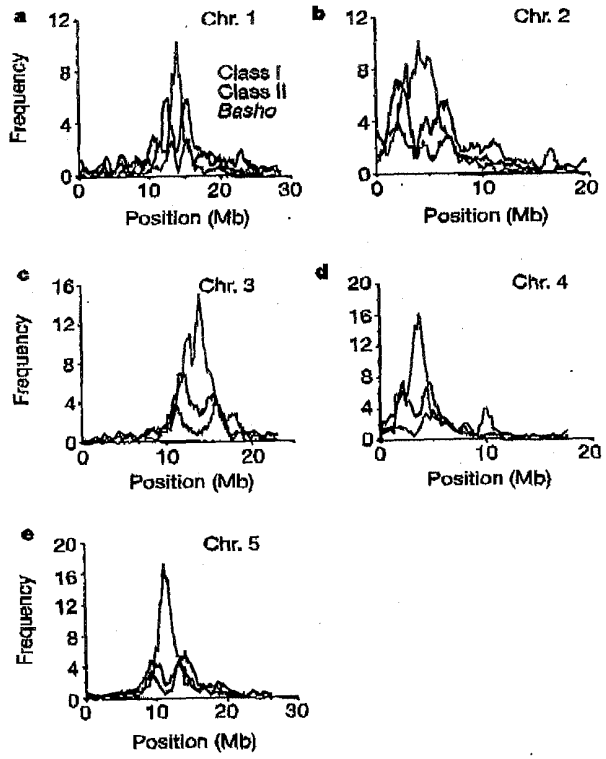

Figure 5 Distribution of class I, II and *Basho* transposons in *Arabidopsis* chromosomes. The frequency of class I retroelements (green), class II DNA transposons (blue) and *Basho* elements (purple) are shown at 100-kb intervals along the five chromosomes (a–e) of *Arabidopsis*.

Membrane transport

Transporters in the plasma and intracellular membranes of *Arabidopsis* are responsible for the acquisition, redistribution and compartmentalization of organic nutrients and inorganic ions, as well as for the efflux of toxic compounds and metabolic end products, energy and signal transduction, and turgor generation. Previous genomic analyses of membrane transport systems in *S. cerevisiae* and *C. elegans* led to the identification of over 100 distinct families of membrane transporters[60,61]. We compared membrane transport processes between *Arabidopsis*, animals, fungi and prokaryotes, and identified over 600 predicted membrane transport systems in *Arabidopsis* (http://www-biology.ucsd.edu/~ipaulsen/transport/), a similar number to that of *C. elegans*

(~700 transporters) and over twofold greater than either *S. cerevisiae* or *E. coli* (~300 transporters).

We compared the transporter complement of *Arabidopsis*, *C. elegans* and *S. cerevisiae* in terms of energy coupling mechanisms (Fig. 7a). Unlike animals, which use a sodium ion P-type ATPase pump to generate an electrochemical gradient across the plasma membrane, plants and fungi use a proton P-type ATPase pump to form a large membrane potential (−250 mV)[62]. Consequently, plant secondary transporters are typically coupled to protons rather than to sodium[63]. Compared with *C. elegans*, *Arabidopsis* has a surprisingly high percentage of primary ATP-dependent transporters (12% and 21% of transporters, respectively), reflecting increased numbers of P-type ATPases involved in metal ion transport and ABC ATPases proposed to be involved in sequestering unusual metabolites and drugs in the vacuole or in other intracellular compartments. These processes may be necessary for pathogen defence and nutrient storage.

About 15% of the transporters in *Arabidopsis* are channel proteins, five times more than in any single-celled organism but half the number in *C. elegans* (Fig. 7b). Almost half of the *Arabidopsis* channel proteins are aquaporins, and *Arabidopsis* has 10-fold more Mfamily major intrinsic protein (MIP) family water channels than any other sequenced organism. This abundance emphasizes the importance of hydraulics in a wide range of plant processes, including sugar and nutrient transport into and out of the vasculature, opening of stomatal apertures, cell elongation and epinastic movements of leaves and stems. Although *Arabidopsis* has a diverse range of metal cation transporters, *C. elegans* has more, many of which function in cell–cell signalling and nerve signal transduction. *Arabidopsis* also possesses transporters for inorganic anions such as phosphate, sulphate, nitrate and chloride, as well as for metal cation channels that serve in signal transduction or cell homeostasis. Compared with other sequenced organisms, *Arabidopsis* has 10-fold more predicted peptide transporters, primarily of the proton-dependent oligopeptide transport (POT) family, emphasizing the importance of peptide transport or indicating that there is broader substrate specificity than previously realized. There are nearly 1,000 *Arabidopsis* genes encoding Ser/Thr protein kinases, suggesting that peptides may have an important role in plant signalling[64].

Virtually no transporters for carboxylates, such as lactate and pyruvate, were identified in the *Arabidopsis* genome. About 12% of the transporters were predicted to be sugar transporters, mostly consisting of paralogues of the MFS family of hexose transporters. Notably, *S. cerevisiae*, *C. elegans* and most prokaryotes use APC family transporters as their principle means of amino-acid transport, but *Arabidopsis* appears to rely primarily on the AAAP family of amino-acid and auxin transporters. More than 10% of the transporters in *Arabidopsis* are homologous to drug efflux pumps; these probably represent transporters involved in the sequestration into vacuoles of xenobiotics, secondary metabolites, and breakdown products of chlorophyll.

Surprisingly, *Arabidopsis* has close homologues of the human ABC TAP transporters of antigenic peptides for presentation to the major histocompatability complex (MHC). In *Arabidopsis*, these transporters may be involved in peptide efflux, or more speculatively, in some form of cell-recognition response. *Arabidopsis* also has 10-fold more members of the multi-drug and toxin extrusion (MATE) family than any other sequenced organism; in bacteria, these transporters function as drug efflux pumps. Curiously, *Arabidopsis* has several homologues of the *Drosophila* RND transporter family Patched protein, which functions in segment polarity, and more than ten homologues of the *Drosophila* ABC family eye pigment transporters. In plants, these are presumably involved in intracellular sequestration of secondary metabolites.

DNA repair and recombination

DNA repair and recombination pathways have many functions in different species such as maintaining genomic integrity, regulating mutation rates, chromosome segregation and recombination, genetic exchange within and between populations, and immune system development. Comparing the *Arabidopsis* genome with other species[65] indicates that *Arabidopsis* has a similar set of DNA repair and recombination (RAR) genes to most other eukaryotes. The pathways represented include photoreactivation, DNA ligation, non-homologous end joining, base excision repair, mismatch excision repair, nucleotide excision repair and many aspects of DNA recombination (Supplementary Information Table 5). The *Arabidopsis* RAR genes include homologues of many DNA repair genes that are defective in different human diseases (for example, hereditary breast cancer and non-polyposis colon cancer, xeroderma pigmentosum and Cockayne's syndrome).

One feature that sets *Arabidopsis* apart from other eukaryotes is the presence of additional homologues of many RAR genes. This is seen for almost every major class of DNA repair, including recombination (four RecA), DNA ligation (four DNA ligase 1), photoreactivation (one class II photolyase and five class I photolyase homologues) and nucleotide excision repair (six RPA1, two RPA2, two Rad25, three TFB1 and four Rad23). This is most striking for genes with probable roles in base excision repair. *Arabidopsis* encodes 16 homologues of DNA base glycosylases (enzymes that

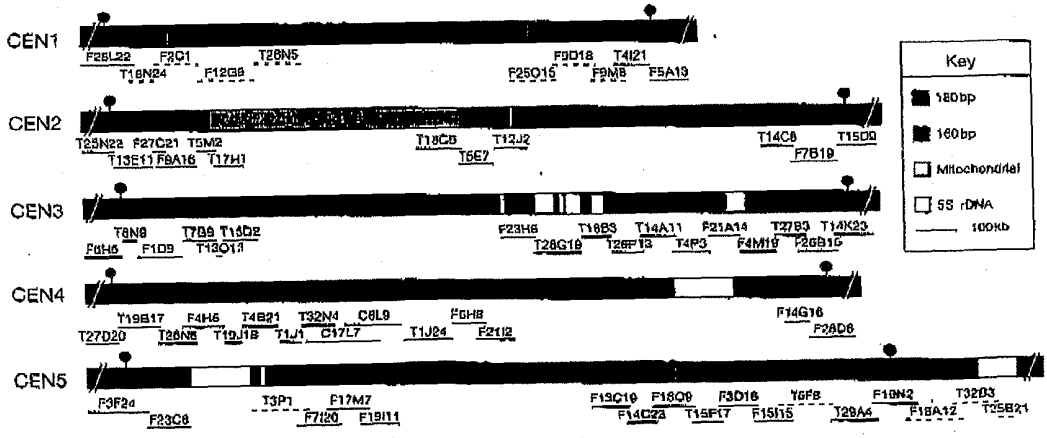

Figure 6 Predicted centromere composition. Genetically defined centromere boundaries are indicated by filled circles; fully and partially assembled BAC sequences are represented by solid and dashed black lines, respectively. Estimates of repeat sizes within the centromeres were derived from consideration of repeat copy number, physical mapping and cytogenetic assays.

recognize abnormal DNA bases and cleave them from the sugar-phosphate backbone)—more than any other species known. This includes several homologues of each of three families of alkylation damage base glycosylases: two of the *S. cerevisiae* MPG; six of the *E. coli* TagI; and two of the *E. coli* AlkA. *Arabidopsis* also encodes three homologues of the apurinic-apyrimidinic (AP) endonuclease Xth. AP endonucleases continue the base excision repair started by glycosylases by cleaving the DNA backbone at abasic sites.

Evolutionary analysis indicates that some of the extra copies of RAR genes in *Arabidopsis* originated through relatively recent gene duplications—because many of the sets of genes are more closely related to each other than to their homologues in any other species. As duplication is frequently accompanied by functional divergence, the duplicate (paralogous) genes may have different repair specificities or may have evolved functions that are outside RAR functions (as is the case for two of the five class I photolyase homologues, which function as blue-light receptors). In most cases, it is not known whether the paralogous gene copies have different functions. The presence of multiple paralogues might also allow functional redundancy or a greater repair or recombination capacity.

The multiplicity of RAR genes in *Arabidopsis* is also partly due to the transfer of genes from the organellar genomes to the nucleus. Repair gene homologues that appear to be of chloroplast origin (Supplementary Information Tables 2 and 5) include the recombination proteins RecA, RecG and SMS, two class I photolyase homologues, Fpg, two MutS2 proteins, and the transcription-repair coupling factor Mfd. Two of these (RecA and Fpg) are involved in RAR functions in the plastid, suggesting that the others may be as well. The finding of an Mfd orthologue of cyanobacterial descent is surprising. In *E. coli*, Mfd couples nucleotide excision repair carried out by UvrABC to transcription, leading to the rapid repair of DNA damage on the transcribed strand of transcribed genes[66] The absence of orthologues of UvrABC in *Arabidopsis* renders the function of Mfd difficult to predict. The presence of Mfd but not UvrABC has been reported for only one other species, a bacterial endosymbiont of the pea aphid.

Other nuclear-encoded *Arabidopsis* DNA repair gene homologues are evolutionarily related to genes from α-Proteobacteria, and thus may be of mitochondrial descent. In particular, the six homologues of the alkyl-base glycosylase TagI appear to be the result of a large expansion in plants after transfer from the mitochondrial genome. Whether any of these TagI homologues function in the repair and maintenance of mitochondrial DNA has not been determined. More detailed phylogenetic analysis may reveal additional *Arabidopsis* RAR genes to be of organellar ancestry.

There are some notable absences of proteins important for RAR in other species, including alkyltransferases, MSH4, RPA3 and many components of TFIIH (TFB2, TFB3, TFB4, CCL1, Kin28). Nevertheless, *Arabidopsis* shows many similarities to the set of DNA repair genes found in other eukaryotes, and therefore offers an experimental system for determining the functions of many of these proteins, in part through characterization of mutants defective in DNA repair[67].

Gene regulation

Eukaryotic gene expression involves many nuclear proteins that modulate chromatin structure, contribute to the basal transcription machinery, or mediate gene regulation in response to developmental, environmental or metabolic cues. As predicted by sequence similarity, more than 3,000 such proteins may be encoded by the *Arabidopsis* genome, suggesting that it has a comparable complexity of gene regulation to other eukaryotes. *Arabidopsis* has an additional level of gene regulation, however, with DNA methylation potentially mediating gene silencing and parental imprinting.

Plants have evolved several variations on chromatin remodelling proteins, such as the family of HD2 histone deacetylases[68]. Although *Arabidopsis* possesses the usual number of SNF2-type chromatin remodelling ATPases, which regulate the expression of nearly all genes, there are significant structural differences between yeast and metazoan SNF2-type genes and their orthologues in *Arabidopsis*. DDM1, a member of the SNF2 superfamily, and MOM1, a gene with similarity to the SNF2 family, are involved in transcriptional gene silencing in *Arabidopsis*. MOM1 has no clear orthologue in fungal or metazoan genomes.

Consistent with its methylated DNA, *Arabidopsis* possesses eight DNA methyltransferases (DMTs). Two of the three types are orthologous to mammalian DMT[69] whereas one, chromomethyltransferase[70], is unique to plants. No DMTs are found in yeast or *C. elegans*, although two DMT-like genes are found in *Drosophila*[71]. *Arabidopsis* also encodes eight proteins with methyl-DNA-binding domains (MBDs). Despite lacking methylated DNA, *Drosophila* encodes four MBD proteins and *C. elegans* has two. These differences in chromatin components are likely to reflect important differences in chromatin-based regulatory control of gene expression in eukaryotes (Supplementary Information Table 6; http://Ag.Arizona.Edu/chromatin/chromatin.html).

The *Arabidopsis* genome encodes transcription machinery for the three nuclear DNA-dependent RNA polymerase systems typical of eukaryotes (Supplementary Information Table 6). Transcription by RNA polymerases II and III appears to involve the same machinery as is used in other eukarotes; however, most transcription factors for RNA polymerase I are not readily identified. Only two polymerase I regulators (other than polymerase subunits and TATA-binding protein) are apparent in *Arabidopsis*, namely homologues of yeast RRN3 and mouse TTF-1. All eukaryotes examined to date have distinct genes for the largest and second largest subunits of polymerase I, II and III. Unexpectedly, *Arabidopsis* has two genes encoding a fourth class of largest subunit and second-largest subunit (Supplementary Information Fig. 5). It will be interesting to determine whether the atypical subunits comprise a polymerase that has a plant-specific function. Four genes encoding single-subunit plastid or mitochondrial RNA polymerases have been identified in *Arabidopsis* (Supplementary Information Table 6). Genes for the bacterial β-, β′- and α-subunits of RNA polymerase are also present, as are homologues of various σ-factors, and these proteins may regulate chloroplast gene expression. Mutations in the *Sde-1* gene, encoding RNA-dependent RNA polymerase (RdRp), lead to defective post-transcriptional gene silencing[72]. We also identified five more closely related RdRp genes.

Our analysis, using both similarity searches and domain matches, has identified 1,709 proteins with significant similarity to known classes of plant transcription factors classified by conserved DNA-binding domains. This analysis used a consistent conservative threshold that probably underestimates the size of families of diverse sequence. This class of protein is the least conserved among all classes of known proteins, showing only 8–23% similarity to transcription factors in other eukaryotes (Fig. 2b). This reduced similarity is due to the absence of certain classes of transcription factors in *Arabidopsis* and large numbers of plant-specific transcription factors. We did not detect any members of several widespread families of transcription factors, such as the REL (Rel-like DNA-binding domain) homology region proteins, nuclear steroid receptors and forkhead-winged helix and POU (Pit-1, Oct- and Unc-8b) domain families of developmental regulators. Conversely, of 29 classes of *Arabidopsis* transcription factors, 16 appear to be unique to plants (Supplementary Information Table 6). Several of these, such as the AP2/EREBP-RAV, NAC and ARF-AUX/IAA families, contain unique DNA-binding domains, whereas others contain plant-specific variants of more widespread domains, such as the DOF and WRKY zinc-finger families and the two-repeat MYB family.

Functional redundancy among members of large families of closely related transcription factors in *Arabidopsis* is a significant potential barrier to their characterization[73]. For example, in the

SHATTERPROOF and SEPALLATA families of MADS box transcription factors, all genes must be defective to produce visible mutant phenotypes[74,75]. These functionally redundant genes are found on the segmental duplications described above. Our analyses, together with the significant sequence similarity found in large families of transcription factors such as the R2R3-repeat MYB and WRKY families, suggest that strategies involving overexpression will be important in determining the functions of members of transcription factor families.

Arabidopsis has two or over three times more transcription factors than identified in *Drosophila*[29] or *C.elegans*[1], respectively. The significantly greater extent of segmental chromosomal and local tandem duplications in the *Arabidopsis* genome generates larger gene families, including transcription factors. The partly overlapping functions defined for a few transcription factors are also likely to be much more widespread, implicating many sequence-related transcription factors in the same cellular processes. Finally, the expanded number of genes involved in metabolism, defence and environmental interaction in *Arabidopsis* (Fig. 2a), which have few counterparts in *Drosophila* and *C. elegans*, all require additional numbers and classes of transcription factors to integrate gene function in response to a vast range of developmental and environmental cues.

Cellular organization

Plant cells differ from animal cells in many features such as plastids, vacuoles, Golgi organization, cytoskeletal arrays, plasmodesmata linking cytoplasms of neighbouring cells, and a rigid polysaccharide-rich extracellular matrix—the cell wall. Because the cell wall maintains the position of a cell relative to its neighbours, both changes in cell shape and organized cell divisions, involving cytoskeleton reorganization and membrane vesicle targeting, have major roles in plant development. Plant cytokinesis is also unique in that the partitioning membrane is formed *de novo* by vesicle fusion. We compared the *Arabidopsis* genome with those of *C. elegans*, *Drosophila* and yeast to glimpse the genetic basis of plant-cell-specific features.

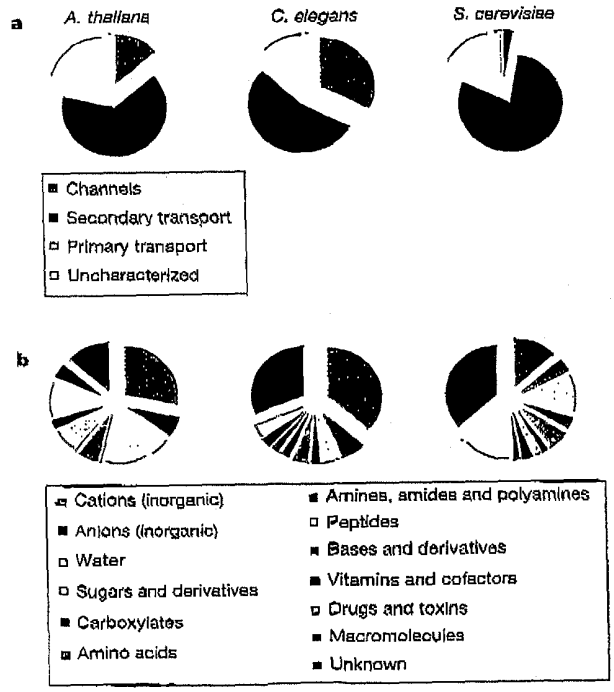

Figure 7 Comparison of the transport capabilities of *Arabidopsis*, *C. elegans* and *S. cerevisiae*. Pie charts show the percentage of transporters in each organism according to bioenergetics (a) and substrate specificity (b).

The principal components of the plant cytoskeleton are microtubules (MTs) and actin filaments (AFs); intermediate filaments (IFs) have not been described in plants. *Arabidopsis* appears to lack genes for cytokeratin or vimentin, the main components of animal IFs, but has several variants of actin, α- and β-tubulin. The *Arabidopsis* genome also encodes homologues of chaperones that mediate the folding of tubulin and actin polypeptides in yeast and animal cells, such as the prefoldin and cytosolic chaperonin complexes and tubulin-folding cofactors. The dynamic stability of MTs and AFs is influenced by MT-associated proteins and actin-binding proteins, respectively, several of which are encoded by *Arabidopsis* genes. These include the MT-severing ATPase katanin, AF-cross-linking/bundling proteins, such as fimbrins and villins, and AF-disassembling proteins, such as profilin and actin-depolymerizing factor/cofilin. The *Arabidopsis* proteome appears to lack homologues of proteins that, in animal cells, link the actin cytoskeleton across the plasma membrane to the extracellular matrix, such as integrin, talin, spectrin, α-actinin, vitronectin or vinculin. This apparent lack of 'anchorage' proteins is consistent with the different composition of the cell wall and with a prominence of cortical MTs at the expense of cortical AFs in plant cells.

Plant-specific cytoskeletal arrays include interphase cortical MTs mediating cell shape, the preprophase band marking the cortical site of cell division, and the phragmoplast assisting in cytokinesis[76]. Although plant cells lack structural counterparts of the yeast spindle pole body and the animal centrosome, *Arabidopsis* has homologues of core components of the MT-nucleating γ-tubulin ring complex, such as γ-tubulin, Spc97/hGCP2 and Spc98/hGCP3. *Arabidopsis* has numerous motor molecules, both kinesins and dyneins with associated dynactin complex proteins, which are presumably involved in the dynamic organization of MTs and in transporting cargo along MT tracks. There are also myosin motors that may be involved in AF-supported organelle trafficking. Essential features of the eukaryotic cytoskeleton appear to be conserved in *Arabidopsis*.

The *Arabidopsis* genome encodes homologues of proteins involved in vesicle budding, including several ARFs and ARF-related small G-proteins, large but not small ARF GEFs (adenosine ribosylation factor on guanine nucleotide exchange factor), adapter proteins, and coat proteins of the COP and non-COP types. *Arabidopsis* also has homologues of proteins involved in vesicle docking and fusion, including SNAP receptors (SNAREs), *N*-ethylmaleimide-sensitive factor (NSF) and Cdc48-related ATPases, accessory proteins such as Sec1 and soluble NSF attachment protein (SNAP), and Rab-type GTPases. The large number of *Arabidopsis* SNAREs can be grouped by sequence similarity to yeast and animal counterparts involved in specific trafficking pathways, and some have been localized to the trans-Golgi and the pre-vacuolar pathway[77]. *Arabidopsis* also has a receptor for retention of proteins in the endoplasmic reticulum, a cargo receptor for transport to the vacuole and several phragmoplastins related to animal dynamin GTPases. Thus, plant cells appear to use the same basic machinery for vesicle trafficking as yeast and animal cells.

Animal cells possess many functionally diverse small G-proteins of the Ras superfamily involved in signal transduction, AF reorganization, vesicle fusion and other processes. Surprisingly, *Arabidopsis* appears to lack genes for G-proteins of the Ras, Rho, Rac and Cdc42 subfamilies but has many Rab-type G-proteins involved in vesicle fusion and several Rop-type G-proteins, one of which has a role in actin organization of the tip-growing pollen tube[78]. The significance of this divergent amplification of different subfamilies of small G-proteins in plants and animals remains to be determined.

Arabidopsis possesses cyclin-dependent kinases (CDKs), including a plant-specific Cdc2b kinase expressed in a cell-cycle-dependent manner, several cyclin subtypes, including a D-type cyclin that

mediates cytokinin-stimulated cell-cycle progression[79], a retinoblastoma-related protein and components of the ubiquitin-dependent proteolytic pathway of cyclin degradation. In yeast and animal cells, chromosome condensation is mediated by condensins, sister chromatids are held together by cohesins such as Scc1, and metaphase–anaphase transition is triggered by separin/Esp1 endopeptidase proteolysis of Scc1 on APC-mediated degradation of its inhibitor, securin/Psd1. Related proteins are encoded by the *Arabidopsis* genome. Thus, the basic machinery of cell-cycle progression, genome duplication and segregation appears to be conserved in plants. By contrast, entry into M phase, M-phase progression and cytokinesis seem to be modified in plant cells. *Arabidopsis* does not appear to have homologues of Cdc25 phosphatase, which activates Cdc2 kinase at the onset of mitosis, or of polo kinase, which regulates M-phase progression in yeast and animals. Conversely, plant-specific mitogen-actived protein (MAP) kinases appear to be involved in cytokinesis.

Cytokinesis partitions the cytoplasm of the dividing cell. Yeast and animal cells expand the membrane from the surface towards the centre in a cleavage process supported by septins and a contractile ring of actin and type II myosin. By contrast, plant cytokinesis starts in the centre of the division plane and progresses laterally. A transient membrane compartment, the cell plate, is formed *de novo* by fusion of Golgi-derived vesicles trafficking along the phragmoplast MTs[80]. Consistent with the unique mode of plant cytokinesis, *Arabidopsis* appears to lack genes for septins and type II myosin. Conversely, cell-plate formation requires a cytokinesis-specific syntaxin that has no close homologue in yeast and animals. Although syntaxin-mediated membrane fusion occurs in animal cytokinesis and cellularization, the vesicles are delivered to the base of the cleavage furrow. Thus, the plant-specific mechanism of cell division is linked to conserved eukaryotic cell-cycle machinery.

Two main conclusions are suggested by this comparative analysis. First, *Arabidopsis* and eukaryotic cells have common features related to intracellular activities, such as vesicle trafficking, cytoskeleton and cell cycle. Second, evolutionarily divergent features, such as organization of the cytoskeleton and cytokinesis, appear to relate to the plant cell wall.

Development

The regulation of development in *Arabidopsis*, as in animals, involves cell–cell communication, hierarchies of transcription factors, and the regulation of chromatin state; however, there is no reason to suppose that the complex multicellular states of plant and animal development have evolved by elaborating the same general processes during the 1.6 billion years since the last common unicellular ancestor of plants and animals[81,82]. Our genome analyses reflect the long, independent evolution of many processes contributing to development in the two kingdoms.

Plants and animals have converged on similar processes of pattern formation, but have used and expanded different transcription factor families as key causal regulators. For example, segmentation in insects and differentiation along the anterior–posterior and limb axes in mammals both involve the spatially specific activation of a series of homeobox gene family members. The pattern of activation is causal in the later differentiation of body and limb axis regions. In plants the pattern of floral whorls (sepals, petals, stamens, carpels) is also established by the spatially specific activation of members of a family of transcription factors, but in this instance the family is the MADS box family. Plants also have homeobox genes and animals have MADS box genes, implying that each lineage invented separately its mechanism of spatial pattern formation, while converging on actions and interactions of transcription factors as the mechanism. Other examples show even greater divergence of plant and animal developmental control. Examples are the AP2/EREBP and NAC families of transcription factors, which have important roles in flower and meristem development; both families are so far found only in plants (Supplementary Information Table 6).

A similar story can be told for cell–cell communication. Plants do not seem to have receptor tyrosine kinases, but the *Arabidopsis* genome has at least 340 genes for receptor Ser/Thr kinases, belonging to many different families, defined by their putative extracellular domains (Supplementary Information Table 7). Several families have members with known functions in cell–cell communication, such as the CLV1 receptor involved in meristem cell signalling, the S-glycoprotein homologues involved in signalling from pollen to stigma in self-incompatible *Brassica* species, and the BRI1 receptor necessary for brassinosteroid signalling[83]. Animals also have receptor Ser/Thr kinases, such as the transforming growth factor-β (TGF-β) receptors, but these act through SMAD proteins that are absent from *Arabidopsis*. The leucine-rich repeat (LRR) family of *Arabidopsis* receptor kinases shares its extracellular domain with many animal and fungal proteins that do not have associated kinase domains, and there are at least 122 *Arabidopsis* genes that code for LRR proteins without a kinase domain. Other *Arabidopsis* receptor kinase families have extracellular domains that are unfamiliar in animals. Thus, evolution is modular, and the plant and animal lineages have expanded different families of receptor kinases for a similar set of developmental processes.

Several *Arabidopsis* genes of developmental importance appear to be derived from a cyanobacteria-like genome (Supplementary Information Table 2), with no close relationship to any animal or fungal protein. One salient example is the family of ethylene receptors; another gene family of apparent chloroplast origin is the phytochromes—light receptors involved in many developmental decisions (see below). Whereas the land plant phytochromes show clear homology to the cyanobacterial light receptors, which are typical prokaryotic histidine kinases, the plant phytochromes are histidine kinase paralogues with Ser/Thr specificity[84]. Similarly to the ethylene receptors, the proteins that act downstream of plant phytochrome signalling are not found in cyanobacteria, and thus it appears that a bacterial light receptor entered the plant genome through horizontal transfer, altered its enzymatic activity, and became linked to a eukaryotic signal transduction pathway. This infusion of genes from a cyanobacterial endosymbiont shows that plants have a richer heritage of ancestral genes than animals, and unique developmental processes that derive from horizontal gene transfer.

Signal transduction

Being generally sessile organisms, plants have to respond to local environmental conditions by changing their physiology or redirecting their growth. Signals from the environment include light and pathogen attack, temperature, water, nutrients, touch and gravity. In addition to local cellular responses, some stimuli are communicated across the plant body, with plant hormones and peptides acting as secondary messengers. Some hormones, such as auxin, are taken up into the cell, whereas others, such as ethylene and brassinosteroids, and the peptide CLV3, act as ligands for receptor kinases on the plasma membrane. No matter where the signal is perceived by the cell, it is transduced to the nucleus, resulting in altered patterns of gene expression.

Comparative genome analysis between *Arabidopsis*, *C. elegans* and *Drosophila* supports the idea that plants have evolved their own pathways of signal transduction[85]. None of the components of the widely adopted signalling pathways found in vertebrates, flies or worms, such as Wingless/Wnt, Hedgehog, Notch/lin12, JAK/STAT, TGF-β/SMADs, receptor tyrosine kinase/Ras or the nuclear steroid hormone receptors, is found in *Arabidopsis*. By contrast, brassinosteroids are ligands of the BRI1 Ser/Thr kinase, a member of the largest recognizable class of transmembrane sensors encoded by 340 receptor-like kinase (RLK) genes in the *Arabidopsis* genome (Supplementary Information Table 7). With a few notable exceptions, such as CLV1, the types of ligands sensed by RLKs are

completely unknown, providing an enormous future challenge for plant biologists. G-protein-coupled receptors (GPCRs)/ seven-transmembrane proteins are an abundant class of proteins in mammalian genomes, instrumental in signal transduction. INTERPRO detected 27 GPCR-related domains in *Arabidopsis* (Supplementary Information Table 1), although there is no direct experimental evidence for these. *Arabidopsis* contains a family of 18 seven-transmembrane proteins of the mildew resistance (MlO) class, several of which are involved in defence responses. Notably, only single Gα (GPA1) and Gβ (AGB1) subunits are found in *Arabidopsis*, both previously known[86].

Although cyclic GMP has been proposed to be involved in signal transduction in *Arabidopsis*[87], a protein containing a guanylate cyclase domain was not identified in our analyses. Nevertheless, cyclic nucleotide-binding domains were detected in various proteins, indicating that cNMPs may have a role in plant signal transduction. Thus, although cNMP-binding domains appear to have been conserved during evolution, cNMP synthesis in *Arabidopsis* may have evolved independently.

We were unable to identify a protein with significant similarity to known Gγ subunits, but recent biochemical studies suggest that a protein with this functional capacity is likely to be present in plant cells (H. Ma, personal communication). Therefore, there is potential for the formation of only a single heterotrimeric G-protein complex; however, its functional interaction with any of the potential GPCR-related proteins remains to be determined.

Modules of cellular signal pathways from bacteria and animals have been combined and new cascades have been innovated in plants. A pertinent example is the response to the gaseous plant hormone ethylene[88]. Ethylene is perceived and its signal transmitted by a family of receptors related to bacterial-type two-component histidine kinases (HKs). In bacteria, yeast and plants, these proteins sense many extracellular signals and function in a His-to-Asp phosphorelay network[89]. In turn, these proteins physically interact with the genetically downstream protein CTR1, a Raf/MAPKKK-related kinase, revealing the juxtaposition of bacterial-type two-component receptors and animal-type MAP kinase cascades. Unlike animals, however, *Arabidopsis* does not seem to have a Ras protein to activate the MAP kinase cascade. MAP kinases are found in abundance in *Arabidopsis*: we identified ~20, a higher number than in any other eukaryote. As potentially counteracting components, we found ~70 putative PP2C protein phosphatases. Although this group is largely uncharacterized functionally, several members are related to ABI1/ABI2, key negative regulators in the signalling pathway for the plant hormone abscisic acid. Additional components of the His-to-Asp phosphorelay system were also found in *Arabidopsis*, including authentic response regulators (ARRs), pseudoresponse regulators (PRRs) and phosphotransfer intermediate protein (HPt)[90]. We found 11 HKs in the proteome (3 new), 16 RRs (2 new) and 8 PRRs (2 new). The biological roles of most ARRs, PRRs and HPts are largly unknown, but several have been found to have diverse functions in plants, including transcriptional activation in response to the plant hormone cytokinin[91], and as components of the circadian clock[92].

Plants seem to have evolved unique signalling pathways by combining a conserved MAP kinase cascade module with new receptor types. In many cases, however, the ligands are unknown. Conversely, some known signalling molecules, such as auxin, are still in search of a receptor. Auxin signalling may represent yet another plant-specific mode of signalling, with protein degradation through the ubiquitin-proteasome pathway preceding altered gene expression. With many *Arabidopsis* genes encoding components of the ubiquitin-proteasome pathway, elimination of negative regulators may be a more widespread phenomenon in plant signalling.

Recognizing and responding to pathogens

Plants are constantly exposed to pests, parasites and pathogens and have evolved many defences. In mammals, polymorphism for parasite recognition encoded in the MHC genes contributes to resistance. In plants, disease resistance (*R*) genes that confer parasite recognition are also extremely polymorphic. This polymorphism has been proposed to restrict parasites, and its absence may explain the breakdown of resistance in crop monocultures[93]. In contrast to MHC genes, plant resistance genes are found at several loci, and the complete genome sequence enables analysis of their complement and structure. Parasite recognition by resistance genes triggers defence mechanisms through various signalling molecules, such as protein kinases and adapter proteins, ion fluxes, reactive oxygen intermediates and nitric oxide. These halt pathogen colonization through transcriptional activation of defence genes and a form of programmed cell death called the hypersensitive response[94]. The *Arabidopsis* genome contains diverse resistance genes distributed at many loci, along with components of signalling pathways, and many other genes whose role in disease resistance has been inferred from mutant phenotypes.

Most resistance genes encode intracellular proteins with a nucleotide-binding (NB) site typical of small G proteins, and carboxy-terminal LRRs[95]. Their amino termini either carry a TIR domain, or a putative coiled coil (CC). There are 85 TIR–NB–LRR resistance genes at 64 loci, and 36 CC–NB–LRR resistance genes at 30 loci. Some NB–LRR resistance genes express neither obvious TIR nor CC domains at their N termini. This potential class is present seven times, at six loci. There are 15 truncated TIR–NB genes that lack an LRR at 10 loci, often adjacent to full TIR–NB–LRR genes. There are also six CC–NB genes, at five loci. These truncated products may function in resistance. Intriguingly, two TIR–NB–LRR genes carry a WRKY domain, found in transcription factors that are implicated in plant defence, and one of these also encodes a protein kinase domain.

Resistance gene evolution may involve duplication and divergence of linked gene families[36]; however, most (46) resistance genes are singletons; 50 are in pairs, 21 are in 7 clusters of 3 family members, with single clusters of 4, 5, 7, 8 and 9 members, respectively. Of the non-singletons, ~60% of pairs are in direct repeats, and ~40% are in inverted repeats. Resistance genes are unevenly distributed between chromosomes, with 49 on chromosome 1; 2 on chromosome 2; 16 on chromosome 3; 28 on chromosome 4; and 55 on chromosome 5.

In other plant species, resistance genes encode both transmembrane receptors for secreted pathogen products and protein kinases, and some other classes are also found. The *Cf* genes in tomato encode extracellular LRRs with a transmembrane domain and short cytoplasmic domain. Mutation in an *Arabidopsis* homologue, *CLAVATA2*, results in enlarged meristems, but to date no resistance function has been assigned to the 30 *Arabidopsis CLV2* homologues. *CLAVATA1*, a transmembrane LRR kinase, is also required for meristem function. *Xa21*, a rice LRR-kinase, confers *Xanthomonas* resistance, and the *Arabidopsis FLS2* LRR kinase confers recognition of flagellin. It has been proposed that CLV1 and CLV2 function as a heterodimer; perhaps this is also true for Xa21, FLS2 and Cf proteins. There are 174 LRR transmembrane kinases in *Arabidopsis*, with only *FLS2* assigned a role in resistance. A unique resistance gene, beet *Hs1pro-1*, which confers nematode resistance, has two *Arabidopsis* homologues.

The tomato Pto Ser/Thr kinase acts as a resistance protein in conjunction with an NB–LRR protein, so similar kinases might do the same for *Arabidopsis* NB–LRR proteins. There are 860 Ser/Thr kinases in the *Arabidopsis* sequence. Fifteen of these share 50% identity over the Pto-aligned region. The Toll pathway in *Drosophila* and mammals regulates innate immune responses through LRR/TIR domain receptors that recognize bacterial lipopolysaccharides[96]. Pto is highly homologous to *Drosophila* PELLE and mammalian IRAK protein kinases that mediate the TIR pathway.

Additional genes have been defined that are required for resistance by our analysis of the genome sequence. The *ndr1* mutation defines a gene required by the CC–NB–LRR gene *RPS2* and *RPM1*. *NDR1* is 1 of 28 *Arabidopsis* genes that are similar both to each other and to the tobacco *HIN1* gene that is transcriptionally induced early during the hypersensitive response. *EDS1* is a gene required for TIR–NB–LRR function, and like *PAD4*, encodes a protein with a putative lipase motif. *EDS1*, *PAD4* and a third gene comprise the *EDS1/PAD4* family. The *NPR1/NIM1/SAI1* gene is required for systemic acquired resistance, and we found five additional *NPR1* homologues. Recessive mutations at both the barley Mlo and *Arabidopsis LSD1* loci confer broad-spectrum resistance and derepress a cell-death program. There are at least 18 Mlo family members that resemble heterotrimeric GPCRs in *Arabidopsis*, and only two *LSD1* homologues.

One of the earliest responses to pathogen recognition is the production of reactive oxygen intermediates. This involves a specialized respiratory burst oxidase protein that transfers an electron across the plasma membrane to make superoxide. *Arabidopsis* encodes eight apparently functional *gp91* homologues, called *Atrboh* genes. Unlike *gp91*, they all carry an ~300 amino-acid N-terminal extension carrying an EF-hand Ca^{2+}-binding domain. In mammals, activation of the respiratory oxidative burst complex in the neutrophil, which includes gp91, requires the action of Rac proteins. As no Rac or Ras proteins are found in *Arabidopsis*, members of the large rop family of G proteins may carry this out. Similarly, we did not detect any *Arabidopsis* homologues of other mammalian respiratory burst oxidase components (p22, p47, p67, p40).

There are no clear homologues of many mammalian defence and cell-death control genes. Although nitric oxide production is involved in plant defence, there is no obvious homologue of nitric oxide synthase. Also absent are apparent homologues of the REL domain transcription factors involved in innate immunity in both *Drosophila* and mammals. We found no similarity to proteins involved in regulating apoptosis in animal cells, such as classical caspases, bcl2/ced9 and baculovirus p35. There are, however, 36 cysteine proteases. There are also eight homologues of a newly defined metacaspase family[97], two of which, along with LSD1, have a clear GATA-type zinc-finger.

Photomorphogenesis and photosynthesis

Because nearly all plants are sessile and most depend on photosynthesis, they have evolved unique ways of responding to light. Light serves as an energy source, as well as a trigger and modulator of complex developmental pathways, including those regulated by the circadian clock. Light is especially important during seedling emergence, where it stimulates chlorophyll production, leaf development, cotyledon expansion, chloroplast biogenesis and the coordinated induction of many nuclear- and chloroplast-encoded genes, while at the same time inhibiting stem growth. The goal of this process, called photomorphogenesis, is the establishment of a body plan that allows the plant to be an efficient photosynthetic machine under varying light conditions[98]. The signal transduction cascade leading to light-induced responses begins with the activation of photoreceptors. Next, the light signal is transduced via positively and negatively acting nuclear and cytoplasmic proteins, causing activation or derepression of nuclear and chloroplast-encoded photosynthetic genes and enabling the plant to establish optimal photoautotrophic growth. Although genetic and biochemical studies have defined many of the components in this process, the genome sequence provides an opportunity to identify comprehensively *Arabidopsis* genes involved in photomorphogenesis and the establishment of photoautotrophic growth. We identified at least 100 candidate genes involved in light perception and signalling, and 139 nuclear-encoded genes that potentially function in photosynthesis.

The roles have been described of only 35 of the 100 candidate photomorphogenic genes (Supplementary Information Table 8). All of the light photoreceptors had been discovered previously, including five red/far-red absorbing phytochromes (PHYA-E), two blue/ultraviolet-A absorbing cryptochromes (CRY1 and CRY2), one blue-absorbing phototropin (NPH1) and one NPH1-like (or NPL1). In contrast, we uncovered many new proteins similar to the photomorphogenesis regulators COP/DET/FUS, PKS1, PIF3, NDPK2, SPA1, FAR1, GIGANTEA, FIN219, HY5, CCA1, ATHB-2, ZEITLUPE, FKF1, LKP1, NPH3 and RPT2.

Both the phytochromes and NPH1 contain chromophores for light sensing coupled to kinase domains for signal transmission. Phytochromes have an N-terminal chromophore-binding domain, two PAS domains, and a C-terminal Ser/Thr kinase domain[99], whereas NPH1 has two LOV domains (members of the PAS domain superfamily) for flavin mononucleotide binding and a C-terminal Ser/Thr kinase domain[100]. PAS domains potentially sense changes in light, redox potential and oxygen energy levels, as well as mediating protein–protein interactions[99,100]. We searched for uncharacterized proteins with the combination of a kinase domain and either a phytochrome chromophore-binding site or PAS domains. Although we found no new phytochrome-like genes, we did identify four predicted proteins that contain PAS and kinase domains (Supplementary Information Fig. 6). These proteins share 80% amino-acid identity, but, unlike NPH1 and NPL1, have only one PAS domain. The combination of potential signal sensing and transmitting domains makes it tempting to speculate that these proteins may be receptors for light or other signals.

Our screen included searches for components of photosynthetic reaction centres and light-harvesting complexes, enzymes involved in CO_2 fixation and enzymes in pigment biosynthesis. We identified 11 core proteins of photosystem I, including the eukaryotic-specific components PsaG and PsaH[101], and 8 photosystem II proteins, including a single member (psbW) of the photosystem II core. We also found 26 proteins similar to the Chlorophyll-a/b binding proteins (8 Lhca and 18 Lhcb). Of the seven subunits of the cytochrome b_6f complex (PetA–D, PetG, PetL, PetM), only one (*PetC*) was found in the nuclear genome, whereas the remainder are probably encoded in the chloroplast. Similarly, of the nine subunits of the chloroplast ATP synthase complex, three are encoded in the nucleus, including the II- , γ- and δ-subunits; the remaining subunits (I, III, IV, α, β, ε) are encoded in the chloroplast[102]. Ten genes were related to the soluble components of the electron transfer chain, including two plastocyanins, five ferredoxins and three ferredoxin/NADP oxidoreductases. Forty genes are predicted to have a role in CO_2 fixation, including all of the enzymes in the Calvin–Benson cycle. For pigment biosynthesis, 16 genes in chlorophyll biosynthesis and 31 genes in carotenoid biosynthesis were found (Supplementary Information Table 8). Our analyses have identified several potential components of the light perception pathway, and have revealed the complex distribution of components of the photosynthetic apparatus between nuclear and plastid genomes.

Metabolism

Arabidopsis is an autotrophic organism that needs only minerals, light, water and air to grow. Consequently, a large proportion of the genome encodes enzymes that support metabolic processes, such as photosynthesis, respiration, intermediary metabolism, mineral acquisition, and the synthesis of lipids, fatty acids, amino acids, nucleotides and cofactors[103]. With respect to these processes, *Arabidopsis* appears to contain a complement of genes similar to those in the photoautotropic cyanobacterium *Synechocystis*[45], but, whereas *Synechocystis* generally has a single gene encoding an enzyme, *Arabidopsis* frequently has many. For example, *Arabidopsis* has at least seven genes for the glycolytic enzyme pyruvate kinase, with an

additional five for pyruvate kinase-like proteins. Whatever the reason for this high level of redundancy, it varies from gene to gene in the same pathway; the 11 enzymes of glycolysis are encoded by up to 51 genes that are present in as few as one or as many as eight copies. Similarly, of the 59 genes encoding proteins involved in glycerolipid metabolism, 39 are represented by more than one gene[104]. Genome duplication and expansion of gene families by tandem duplication have contributed to this diversity.

This high degree of apparent structural redundancy does not necessarily imply functional redundancy. For instance, although there are seven genes for serine hydroxymethyltransferase, a mutation in the gene for the mitochondrial form completely blocks the photorespiratory pathway[105]. Although there are 12 genes for cellulose synthase, mutations in at least 2 of the 12 confer distinct phenotypes because of tissue-specific gene expression[106].

The metabolome of *Arabidopsis* differs from that of cyanobacteria, or of any other organism sequenced to date, by the presence of many genes encoding enzymes for pathways that are unique to vascular plants. In particular, although relatively little is known about the enzymology of cell-wall metabolism, more than 420 genes could be assigned probable roles in pathways responsible for the synthesis and modification of cell-wall polymers. Twelve genes encode cellulose synthase, and 29 other genes encode 6 families of structurally related enzymes thought to synthesize other major polysaccharides[106]. Roughly 52 genes encode polygalacturonases, 20 encode pectate lyases and 79 encode pectin esterases, indicating a massive investment in modifying pectin. Similarly, the presence of 39 β-1,3-glucanases, 20 endoxyloglucan transglycosylases, 50 cellulases and other hydrolases, and 23 expansins reflects the importance of wall remodelling during growth of plant cells. Excluding ascorbate and glutathione peroxidases, there are 69 genes with significant similarity to known peroxidases and 15 laccases (diphenol oxidases). Their presence in such abundance indicates the importance of oxidative processes in the synthesis of lignin, suberin and other cell-wall polymers. The high degree of apparent redundancy in the genes for cell-wall metabolism might reflect differences in substrate specificity by some of the enzymes.

The high degree of apparent redundancy in the genes for cell wall metabolism might reflect differences in substrate specificity by some of the enzymes. It is already known that cell types have different wall compositions, which may require that the relevant enzymes be subject to cell-type-specific transcriptional regulation. Of the 40 or so cell types that plants make, almost all can be identified by unique features of their cell wall[107]. A large number of genes involved in wall metabolism have yet to be defined. Although more than 60 genes for glycosyltransferases can be found in the genome sequence, most of these are probably involved in protein glycosylation or metabolite catabolism and do not seem to be adequate to account for the polysaccharide complexity of the wall. For instance, at least 21 enzymes are required just to produce the linkages of the pectic polysaccharide RGII, and none of these enzymes has been identified at present. Thus, if these and related enzymes involved in the synthesis of other cell-wall polymers are also represented by multiple genes, a substantial number of the genes of currently unknown function may be involved in cell-wall metabolism.

Higher plants collectively synthesize more than 100,000 secondary metabolites. Because flowering plants are thought to have similar numbers of genes, it is apparent that a great deal of enzyme creation took place during the evolution of higher plants. An important factor in the rapid evolution of metabolic complexity is the large family of cytochrome P450s that are evident in *Arabidopsis* (Supplementary Information Table 1). These enzymes represent a superfamily of haem-containing proteins, most of which catalyse NADPH- and O_2-dependent hydroxylation reactions. Plant P450s participate in myriad biochemical pathways including those devoted to the synthesis of plant products, such as phenylpropanoids, alkaloids, terpenoids, lipids, cyanogenic glycosides and glucosinolates, and plant growth regulators, such as gibberellins, jasmonic acid and brassinosteroids. Whereas *Arabidopsis* has ~286 P450 genes, *Drosophila* has 94, *C. elegans* has 73 and yeast has only 3. This low number in yeast indicates that there are few reactions of basic metabolism that are catalysed by P450s. It seems likely that many animal P450s are involved in detoxification of compounds from food plant sources. The role of endogenous enzymes is poorly understood; only a few dozen P450 enzymes from plants have been characterized to any extent. The discrepancy between the number of known P450-catalysed reactions and the number of genes suggests that *Arabidopsis* produces a relatively large number of metabolites that have yet to be identified.

In addition to the large number of cytochrome P450s, *Arabidopsis* has many other genes that suggest the existence of pathways or processes that are not currently known. For instance, the presence of 19 genes with similarity to anthranilate *N*-hydroxycinnamoyl/benzoyl transferase is currently inexplicable. This enzyme is involved in the synthesis of dianthramide phytoalexins in Caryophyllaceae and Gramineae. No phytoalexins of this class have been described in *Arabidopsis* as yet. Similarly, the presence of 12 genes with sequence similarity to the berberine bridge enzyme, ((*S*)-reticuline:oxygen oxidoreductase (methylene-bridge-forming); EC 1.5.3.9), and 13 genes with similarity to tropinone reductase, suggests that *Arabidopsis* may have the ability to produce alkaloids. In other plants, the berberine bridge enzyme transforms reticuline into scoulerine, a biosynthetic precursor to a multitude of species-specific protopine, protoberberine and benzophenanthridine alkaloids. The discovery of these and many other intriguing genes in the *Arabidopsis* genome has created a wealth of new opportunities to understand the metabolic and structural diversity of higher plants.

Concluding remarks

The twentieth century began with the rediscovery of Mendel's rules of inheritance in pea[108], and it ends with the elucidation of the complete genetic complement of a model plant, *Arabidopsis*. The analysis of the completed sequence of a flowering plant reported here provides insights into the genetic basis of the similarities and differences of diverse multicellular organisms. It also creates the potential for direct and efficient access to a much deeper understanding of plant development and environmental responses, and permits the structure and dynamics of plant genomes to be assessed and understood.

Arabidopsis, *C. elegans* and *Drosophila* have a similar range of 11,000–15,000 different types of proteins, suggesting this is the minimal complexity required by extremely diverse multicellular eukaryotes to execute development and respond to their environment. We account for the larger number of gene copies in *Arabidopsis* compared with these other sequenced eukaryotes with two possible explanations. First, independent amplification of individual genes has generated tandem and dispersed gene families to a greater extent in *Arabidopsis*, and unequal crossing over may be the predominant mechanism involved. Second, ancestral duplication of the entire genome and subsequent rearrangements have resulted in segmental duplications. The pattern of these duplications suggests an ancient polyploidy event, and mutant analysis indicates that at least some of the many duplicate genes are functionally redundant. Their occurrence in a functionally diploid genetic model came as a surprise, and is reminiscent of the situation in maize, an ancient segmental allotetraploid. The remarkable degree of genome plasticity revealed in the large-scale duplications may be needed to provide new functions, as alternative promoters and alternative splicing appear to be less widely used in plants than they are in animals. Apart from duplicated segments, the overall chromosome structure of *Arabidopsis* closely resembles that of *Drosophila*; transposons and other repetitive sequences are concentrated in the heterochromatic regions surrounding the centromere,

whereas the euchromatic arms are largely devoid of repetitive sequences. Conversely, most protein-coding genes reside in the euchromatin, although a number of expressed genes have been identified in centromeric regions. Finally, *Arabidopsis* is the first methylated eukaryotic genome to be sequenced, and will be invaluable in the study of epigenetic inheritance and gene regulation.

Unlike most animals, plants generally do not move, they can perpetuate indefinitely, they reproduce through an extended haploid phase, and they synthesize all their metabolites. Our comparison of *Arabidopsis*, bacterial, fungal and animal genomes starts to define the genetic basis for these differences between plants and other life forms. Basic intracellular processes, such as translation or vesicle trafficking, appear to be conserved across kingdoms, reflecting a common eukaryotic heritage. More elaborate intercellular processes, including physiology and development, use different sets of components. For example, membrane channels, transporters and signalling components are very different in plants and animals, and the large number of transcription factors unique to plants contrasts with the conservation of many chromatin proteins across the three eukaryotic kingdoms. Unexpected differences between seemingly similar processes include the absence of intracellular regulators of cell division (Cdc25) and apoptosis (Bcl-2). On the other hand, DNA repair appears more highly conserved between plants and mammals than within the animal kingdom, perhaps reflecting common factors such as DNA methylation. Our analysis also shows that many genes of the endosymbiotic ancestor of the plastid have been transferred to the nucleus, and the products of this rich prokaryotic heritage contribute to diverse functions such as photoautotrophic growth and signalling.

The sequence reported here changes the fundamental nature of plant genetic analysis. Forward genetics is greatly simplified as mutations are more conveniently isolated molecularly, but at the same time extensive gene duplications mean that functional redundancy must be taken into account. At a biochemical level, the specificity conferred by nucleotide sequence, and the completeness of the survey allow complex mixtures of RNA and protein to be resolved into their individual components using micro-arrays and mass spectrometry. This specificity can also be used in the parallel analysis of genome-wide polymorphisms and quantitative traits in natural populations[109]. Looking ahead, the challenge of determining the function of the large set of predicted genes, many of which are plant-specific, is now a clear priority, and multinational programs have been initiated to accomplish this goal using site-selected mutagenesis among the the necessary tools[110]. Finally, productive paths of crop improvement, based on enhanced knowledge of *Arabidopsis* gene function, will help meet the challenge of sustaining our food supply in the coming years.

Note added in proof: at the time of publication 17 centromeric BACs and 5 sequence gaps in chromosome arms are being sequenced. □

Methods

The three centres used similar annotation approaches involving in silico gene-finding methods, comparison to EST and protein databases, and manual reconciliation of that data. Gene finding involved three steps: (1) analysis of BAC sequences using a computational gene finder; (2) alignment of the sequence to the protein and EST databases; (3) assignment of functions to each of the genes. Genscan[111], GeneMark.HMM[112], Xgrail[113] Genefinder (P. Green, unpublished software) and GlimmerA[114] were used to analyse BAC sequences. All of these systems were specially trained for *Arabidopsis* genes. Splice sites were predicted using NetGene2[115], Splice Predictor[116] and GeneSplicer (M. Pertea and S. Salzberg, unpublished software). For the second step, BACs were aligned to ESTs and to the *Arabidopsis* gene index[117] using programs such as DDS/GAP2[118] or BLASTN[119]. Segmental duplications were analysed and displayed using a modified version of DIALIGN2 (ref. 120).

Received 20 October; accepted 15 November 2000.

1. The *C. elegans* Sequencing Consortium. Sequence and analysis of the genome of *C. elegans*. *Science* 282, 2012–2018 (1998).
2. Adams, M. D. The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195 (2000).
3. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* 282, 662–665 (1998).
4. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402, 761–768 (1999).
5. Mayer, K. *et al.* Sequence and analysis of chromsome 4 of the plant *Arabidopsis thaliana*. *Nature* 402, 769–777 (1999).
6. Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* 408, 816–820 (2000).
7. Salanoubat, M. *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* 408, 820–822 (2000).
8. Tabata, S. *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* 408, 820–822 (2000).
9. Choi, S. D., Creelman, R., Mullet, J. & Wing, R. A. Construction and characterisation of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* 2, 17–20 (1995).
10. Mozo, T., Fischer, S., Shizuya, H. & Altmann, T. Construction and characterization of the IGF *Arabidopsis* BAC library. *Mol. Gen. Genet.* 258, 562–570 (1998).
11. Lui, Y. -G., Mitsukawa, N., Vazquez-Tello, A. & Whittier, R. F. Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* 7, 351–358 (1995).
12. Lui, Y. -G. *et al.* Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc. Natl Acad. Sci. USA* 96, 6535–6540 (1999).
13. Marra, M. *et al.* A map or sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* 22, 265–270 (1999).
14. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* 22, 271–275 (1999).
15. Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1. 6 Mb regions covered by twenty physically assigned P1 clones. *DNA Res.* 4, 215–230 (1997).
16. Bent, E., Johnson, S. & Bancroft, I. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. *Plant J.* 13, 849–855 (1998).
17. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* 286, 2468–2474 (1999).
18. Meyerowitz, E. M. & Somerville, C. R. *Arabidopsis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1994)
19. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964 (1997).
20. Pavy, N. *et al.* Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *BioInformatics* 15, 887–900 (1999).
21. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* 387 (Suppl.) 7–65 (1997).
22. Frishman, D. *et al.* Functional and structural genomics using PEDANT. *BioInformatics* (in the press).
23. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1462 (1997).
24. Kotani, H. & Tabata, S. Lessons from the sequencing of the genome of a unicellular cyanobacterium, *Synechocystis* SP. PCC6803. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49, 151–171 (1998).
25. Apweiler, R. *et al.* *INTERPRO* (http://www.ebi.ac.uk/interpro/). Collaborative Computer Project 11 Newsletter no. 10 (Cambridge, 2000).
26. Bent, A. F. *et al.* *RPS2* of *Arabidopsis thaliana* a leucine-rich repeat class of plant disease resistance genes. *Science* 265, 1856–1860 (1994).
27. Skowyra, D. *et al.* F box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* 91, 209–219 (1997).
28. Jonzeiro, C. A. P. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* 102, 549–552 (2000).
29. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* 287, 2204–2215 (2000).
30. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* 27, 2369–2376 (1999).
31. Blanc, G. *et al.* Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* 12, 1093–1102 (2000).
32. Wendel, J. F. Genome evolution in polyploids. *Plant Mol. Biol.* 42, 225–249 (2000).
33. Gaut, B. S. & Doebley, J. F. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA* 94, 6809–6814 (1997).
34. Ku, H. -M., Vision, T., Liu, J. & Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA* 97, 9121–9126 (2000).
35. Noel, L. *et al.* Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. *Plant Cell* 11, 2099–2111 (1999).
36. Ellis, J., Dodds, P. & Pryor, T. Structure, function, and evolution of plant disease resistance genes. *Trends Plant Sci.* 3, 278–284 (2000).
37. Tanksley, S. D. *et al.* High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132, 1141–1160 (1992).
38. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Grasses, line up and form a circle. *Curr. Biol.* 5, 737–739 (1995).
39. Acarkan, A., Rossberg, M., Koch, M. & Schmidt, R. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. *Plant J.* 23, 55–62 (2000).
40. Cavell, A., Lydiate, D., Parkin, I., Dean, C. & Trick, M. A 30 centimorgan segment of *Arabidopsis thaliana* chromosome 4 has six collinear homologues within the *Brassica napus* genome. *Genome* 41, 62–69 (1998).
41. O'Neill, C. & Bancroft, I. Comparative physical mapping of segments of the genome of *Brassica oleracea* var *alboglabra* that are homologous to sequenced regions of the chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* 23, 233–243 (2000).
42. Wolfe, K. H., Gouy, M., Yang, Y. -W., Sharp, P. M. & Li, W. -H. Date of the monocot-dicot divergence estimated from the chloroplast DNA sequence data. *Proc. Natl Acad. Sci. USA* 86, 6201–6205 (1989).
43. van Dodeweerd, A. -M. *et al.* Identification and analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome* 42, 887–892 (1999)
44. Mayer, K. Sequence level analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*, *Genome Res.* (submitted).
45. Sato, S. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* 6, 283–290 (1999).
46. Unseld, M., Marienfeld, J., Brandt, P. & Brennicke, A. The mitochondrial genome in *Arabidopsis*

thaliana contains 57 genes in 366,924 nucleotides. *Nature Genet.* **15**, 57–61 (1997).
47. Palmer, J. D. *et al.* Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl Acad. Sci. USA* **97**, 6960–6966 (2000).
48. Le, Q.-H. *et al.* Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **97**, 7376–7381 (2000).
49. Yu, Z., Wright, S. & Bureau, T. *Mutator*-like elements (MULEs) in *Arabidopsis thaliana*: Structure, diversity and evolution. *Genetics* (in the press).
50. Feschotte, C. & Mouches, C. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.* **17**, 730–737 (2000).
51. Martienssen, R. Transposons, DNA methylation and gene control. *Trends Genet.* **14**, 263–264 (1998).
52. Singer, T., Yordan, C. & Martienssen, R. Robertson's *Mutator* transposons in *Arabidopsis* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation (DDM1)*. *Genes Dev.* (in the press).
53. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
54. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9**, 273–282 (1996).
55. Fransz, P. *et al.* Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J.* **13**, 867–876 (1998).
56. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53**, 127–136 (1988).
57. Round, E. K., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7**, 1045–1053 (1997).
58. The CSHL/WUGSC/PEB *Arabidopsis* Sequencing Consortium. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100**, 377–386 (2000).
59. Fransz, P. F. *et al.* Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: Structural organization of heterochromatic knob and centromere region. *Cell* **100**, 367–376 (2000).
60. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301**, 75–101 (2000).
61. Paulsen, I. T., Sliwinski, M. K., Nelissen, B., Goffeau, A. & Saier, M. H. Jr Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.* **430**, 116–125 (1998).
62. Hirsch, R. E., Lewis, B. D, Spalding, E. P. & Sussman, M. R. A role for the AKT1 potassium channel in plant nutrition. *Science* **280**, 918–921 (1998).
63. Slayman, C. L. & Slayman, C. W. Depolarization of the plasma membrane of Neurospora during active transport of glucose: evidence for a proton-dependent cotransport system. *Proc. Natl Acad. Sci. USA* **71**, 1935–1939 (1974).
64. Ryan, C. A. & Pearce, G. Systemin: a polypeptide signal for plant defensive genes. *Annu. Rev. Cell. Dev. Biol.* **14**, 1–17 (1998).
65. Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435**, 171–213 (1999).
66. Selby, C. P. & Sancar, A. Structure and function of transcription-repair coupling factor. Structural domains and binding properties. *J. Biol. Chem.* **270**, 4882–4889 (1995).
67. Britt, A. B. Molecular genetics of DNA repair in higher plants. *Trends Plant Sci.* **4**, 20–25 (1999).
68. Dangl, M. Response to Aravind, L. & Koonin, E. V. Second Family of Histone Deacetylases. *Science* **280**, 1167 (1998).
69. Cao, X. *et al.* Conserved plant genes with similarity to mammalian de novo DNA methyltransferases. *Proc. Natl Acad. Sci. USA* **97**, 4979–4984 (2000).
70. Henikoff, S. & Comai, L. A DNA methyltransferase homologue with a chromodomain exists in multiple polymorphic forms in *Arabidopsis*. *Genetics* **149**, 307–318 (1998).
71. Hung, M. -S. *et al. Drosophila* proteins related to vertebrate DNA (5-cytosine) methyltransferases. *Proc. Natl Acad. Sci. USA* **96**, 11940–11945 (1999).
72. Dalmay, T., Hamilton, A. J., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent-RNA polymerase in *Arabidopsis* is required for post transcriptional gene silencing mediated by a transgene but not by a virus—the truth. *Cell* **101**, 543–553 (2000).
73. Riechmann, J. L. & Ratcliffe, O. J. A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **3**, 423–434 (2000).
74. Liljegren, S. J. *et al.* SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404**, 766–770 (2000).
75. Pelaz, S. *et al.* B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* **405**, 200–203 (2000).
76. Canaday, J., Stoppin-Mellet, V., Mutterer, J., Lambert, A. M. & Schmit, A. C. Higher plant cells: gamma-tubulin and microtubule nucleation in the absence of centrosomes. *Microsc. Res. Technol.* **49**, 487–495 (2000).
77. Bassham, D. C. & Raikhel, N. V. Unique features of the plant vacuolar sorting machinery. *Curr. Opin. Cell Biol.* **12**, 491–495 (2000).
78. Zheng, Z. L. & Yang, Z. The Rop GTPase switch turns on polar growth in pollen. *Trends Plant Sci.* **5**, 298–303 (2000).
79. den Boer, B. G. & Murray, J. A. Triggering the cell cycle in plants. *Trends Cell Biol.* **10**, 245–250 (2000).
80. Heese, M., Mayer, U. & Jurgens, G. Cytokinesis in flowering plants: cellular process and developmental integration. *Curr. Opin. Plant Biol.* **1**, 486–491 (1998).
81. Meyerowitz, E. M. Plants, animals, and the logic of development. *Trends Genet.* **15**, M65–M68 (1999).
82. Wang, D. Y. C. *et al.* Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Bio.* **266**, 163–171 (1999).
83. Torii, K. Receptor kinase activation and signal transduction in plants: an emerging picture. *Curr. Opin. Plant Biol.* **3**, 362–367 (2000).
84. Yeh, K. C. & Lagarias, J. C. Eukaryotic phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl Acad. Sci. USA* **95**, 13976–13981 (1998).
85. McCarty, D. R. & Chory, J. Conservation and innovation in plant signaling pathways. *Cell* **103**, 201–211 (2000).
86. Weiss, C. A., Garnaat, C., Mukai, K., Hu, Y. & Ma, H. Molecular cloning of cDNAs from maize and *Arabidopsis* encoding a G protein beta subunit. *Proc. Natl Acad. Sci. USA* **91**, 9554–9558 (1994).
87. Bowler, C. *et al.* Cyclic GMP and calcium mediate phytochrome phototransduction. *Cell* **77**, 73–81 (1994).
88. Stepanova, A. & Ecker, J. R. Ethylene signaling: from mutants to molecules. *Curr. Opin. Plant Biol.* **3**, 353–360 (2000).
89. Urao, T., Yamaguchi-Shinozaki, K. & Shinozaki, K. Two-component systems in plant signal transduction. *Trends Plant Sci.* **5**, 67–74 (2000).
90. Makino, S. *et al.* Genes encoding pseudo-response regulators: Insight into His-to-Asp phosphorelay and circadian rhythm in *Arabidopsis thaliana*. *Plant Cell Physiol.* **41**, 791–803 (2000).
91. D'Agostino, I. B. & Kieber, J. J. Phosphorelay signal transduction: the emerging family of plant response regulators. *Trends Biol. Sci.* **24**, 452–456 (1999).
92. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene TOC1, an autoregulatory response regulator homologue. *Science* **289**, 768–771 (2000).
93. Stahl, E. A. & Bishop, J. G. Plant-Pathogen arms races at the molecular level. *Curr. Opin. Plant Biol.* **3**, 299–304 (2000).
94. McDowell, J. M. & Dangl, J. L. Signal transduction in the plant innate immune response. *Trends Biochem. Sci.* **25**, 79–82 (2000).
95. Van der Biezen, E. A. & Jones, J. D. Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci.* **23**, 454–456 (1998).
96. Belvin, M. P. & Anderson, K. V. A conserved signaling pathway: the *Drosophila* toll-dorsal pathway. *Annu. Rev. Cell. Dev. Biol.* **12**, 393–416 (1996).
97. Uren, A. G. *et al.* Identification of paracaspases and metacaspases: Two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* **6**, 961–967 (2000).
98. Fankhauser, C. & Chory, J. Light control of plant development. *Annu. Rev. Cell. Dev. Biol.* **13**, 203–229 (1997).
99. Briggs, W. R. & Huala, E. Blue-light photoreceptors in higher plants. *Annu. Rev. Cell. Dev. Biol.* **15**, 33–62 (1999).
100. Christie, J. M., Salomon, M., Nozue, K., Wada, M. & Briggs, W. R. LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): binding sites for the chromophore flavin mononucleotide. *Proc. Natl Acad. Sci. USA* **96**, 8779–8783 (1999).
101. Golbeck, J. H. Structure and function of photosystem I. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43**, 293–324 (1992).
102. Maier, R. M., Neckermann, K., Igloi, G. L. & Kossel, H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* **251**, 614–28 (1995).
103. Buchanan, B. B., Gruissem, W. & Jones, R. L. In *Biochemistry and Molecular Biology of Plants* 1367 (Am. Soc. Plant Physiol., Rockville, Maryland, 2000).
104. Mekhedov, S., Martínez de Ilárduya, O. & Ohlrogge, J. Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiol.* **122**, 389–401 (2000).
105. Somerville, C. R., & Ogren, W. L. Photorespiration deficient mutants of *Arabidopsis thaliana* lacking mitochrondrial serine transhydroxymethylase activity. *Plant Physiol.* **67**, 666–671 (1981).
106. Richmond, T., & Somerville, C. R. The cellulose synthase superfamily. *Plant Physiol* **124**, 495–499 (1999).
107. Carpita, N. Vergara C: A recipe for cellulose. *Science* **279**, 672–673 (1998).
108. De Vries, H. Sur la loi de disjonction des hybrides. *C. R. Acad. Sci. Paris* **130**, 845–847 (1900).
109. Alonso-Blanco, C. & Koornneef, M. Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* **5**, 1360–1385 (1999).
110. Chory, J. Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiology* **123**, 423–425 (2000).
111. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
112. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
113. Uberbacher, E. C. & Mural, R. J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88**, 11261–11265 (1991).
114. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
115. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24**, 3439–3452 (1996).
116. Brendel, V. & Kleffe, J. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.* **26**, 4748–4757 (1998).
117. Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28**, 141–145 (2000).
118. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37–45 (1997).
119. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
120. Morgenstern, B. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218 (1999).
121. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
122. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This work was supported by the National Science Foundation (NSF) Cooperative Agreements (funded by the NSF, the US Department of Agriculture (USDA) and the US Department of Energy (DOE)), the Kazusa DNA Research Institute Foundation, and by the European Commission. Additional support from the USDA, Ministère de la Recherche, GSF-Forschungszentrum f. Umwelt u. Gesundheit, BMBF (Bundesministerium f. Bildung, Forschung und Technologie), the BBSRC (Biotechnology and Biological

Sciences Research Council) and the Plant Research International, Wageningen, is also gratefully acknowledged. The authors wish to thank E. Magnien, D. Nasser and J. D. Watson for their continual support and encouragement.

Correspondence and requests for materials should be addressed to The Arabidopsis Genome Initiative (e-mail: genomeanalysis@tigr.org or genomeanalysis@gsf.de).

The Arabidopsis Genome Initiative

Three groups contributed to the work reported here. The Genome Sequencing groups, arranged here in order of sequence contribution, sequenced and annotated assigned chromosomal regions. The Genome Analysis group carried out the analyses described. The Contributing Authors interpreted the genome analyses, incorporating other data and analyses, with respect to selected biological topics.

Genome Sequencing Groups

Samir Kaul, Hean L. Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J. Tallon, Tamara Feldblyum, William Nierman, Maria-Ines Benito, Xiaoying Lin, Christopher D. Town, J. Craig Venter & Claire M. Fraser

The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Satoshi Tabata, Yasukazu Nakamura, Takakazu Kaneko, Shusei Sato, Erika Asamizu, Tomohiko Kato, Hirokazu Kotani & Shigemi Sasamoto

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan

Joseph R. Ecker[1]*†, Athanasios Theologis[2]*, Nancy A. Federspiel[3]*†, Curtis J. Palm[3], Brian J. Osborne[2], Paul Shinn[1], Aaron B. Conway[3], Valentina S. Vysotskaia[2], Ken Dewar[1], Lane Conn[3], Catherine A. Lenz[2], Christopher J. Kim[1], Nancy F. Hansen[3], Shirley X. Liu[2], Eugen Buehler[1], Hootan Altafi[3], Hitomi Sakano[2], Patrick Dunn[1], Bao Lam[3], Paul K. Pham[2], Qimin Chao[1], Michelle Nguyen[3], Guixia Yu[2], Huaming Chen[1], Audrey Southwick[3], Jeong Mi Lee[2], Molly Miranda[3], Mitsue J. Toriumi[2] & Ronald W. Davis[3]

*1, Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104 USA; 2, Plant Gene Expression Center/USDA-U.C.Berkeley, 800 Buchanan Street, Albany, California 94710, USA; 3, Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA. * These authors contributed equally to this work. † Present addresses: The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA (J.R.E.); Exelixis, Inc., 170 Harborway, P.O. Box 511, South San Francisco, California 94083-0511, USA (N.A.F)*

European Union Chromosome 4 and 5 Sequencing Consortium: R. Wambutt[1], G. Murphy[2], A. Düsterhöft[3], W. Stiekema[4], T. Pohl[5], K.-D. Entian[6], N. Terryn[7] & G. Volckaert[8]

1, AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany; 2, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 3, QIAGEN GmbH, Max-Volmer-Str. 4, D-40724 Hilden, Germany; 4, Greenomics, Plant Research International, Droevendaalsesleeg 1, NL 6700, AA Wageningen, The Netherlands; 5, GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany; 6, SRD GmbH, Oberurseler Str. 43, Oberursel 61440, Germany; 7, Department for Plant Genetics, (VIB), University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium; 8, Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium

European Union Chromosome 3 Sequencing Consortium: M. Salanoubat[1], N. Choisne[1], M. Rieger[2], W. Ansorge[3], M. Unseld[4], B. Fartmann[5], G. Valle[6], F. Artiguenave[1], J. Weissenbach[1] & F. Quetier[1]

1, Genoscope and CNRS FRE2231, 2 rue G. Crémieux, 91057 Evry Cedex, France; 2, Genotype GmbH Angelhofweg 39, D-69259 Wilhemlsfeld, Germany; 3, European Molecular Biology Laboratory, Biochemical Instrumentation Program, Meyerhofstr. 1, D-69117 Heidelberg, Germany; 4, LION Bioscience AG, Im Neuenheimer Feld 515-517, 69120 Heidelberg, Germany; 5, MWG-Biotech AG, Anzinger Strasse 7a, 85560 Ebersberg, Germany; 6, CRIBI, Università di Padova, via G. Colombo 3, Padova 35131, Italy

The Cold Spring Harbor and Washington University Genome Sequencing Center Consortium: Richard K. Wilson[1], Melissa de la Bastide[2], M. Sekhon[1], Emily Huang[2], Lori Spiegel[2], Lidia Gnoj[2], K. Pepin[1], J. Murray[1], D. Johnson[1], Kristina Habermann[2], Nellay Dedhia[2], Larry Parnell[2], Raymond Preston[2], L. Hillier[1], Ellson Chen[3], M. Marra[2], Robert Martienssen[4] & W. Richard McCombie[2]

1, Washington University Genome Sequencing Center, Washington University in St Louis School of Medicine, 4444 Forest Park Blvd., St. Louis, Missouri 63108 USA; 2, Lita Annenberg Hazen Genome Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 3, Celera Genomics, 850 Lincoln Center Drive, Foster City, California 94494, USA; 4, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Genome Analysis Group

Klaus Mayer[1]*, Owen White[2]*, Michael Bevan[3], Kai Lemcke[1], Todd H. Creasy[2], Cord Bielke[2], Brian Haas[1], Dirk Haase[1], Rama Maltí[2], Stephen Rudd[1], Jeremy Peterson[2], Heiko Schoof[1], Dimitrij Frishman[1], Burkhard Morgenstern[1], Paulo Zaccaria[1], Maria Ermolaeva[2], Mihaela Pertea[2], John Quackenbush[2], Natalia Volfovsky[2], Dongying Wu[2], Todd M. Lowe[4], Steven L. Salzberg[2] & Hans-Werner Mewes[1]

*1, GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany; 2, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 3, Molecular Genetics Deartment, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 4, Dept Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA. * These authors contributed equally to this work*

Contributing Authors

Comparative analysis of the genomes of *A. thaliana* accessions. S. Rounsley, D. Bush, S. Subramaniam, I. Levin & S. Norris

Cereon Genomics LLC, 45 Sidney St, Cambridge, Massachussetts 02139, USA

Comparative analysis of the genomes of *A. thaliana* and other genera. R. Schmidt[1], A. Acarkan[1] & I. Bancroft[2]

1, Max-Delbrück-Laboratorium in der Max-Planck-Gesellschaft, Carl-von-Linné-Weg 10, 50829 Cologne, Germany; 2, Brassicas and Oilseeds Research Department, John Innes Centre, Norwich NR4 7UJ, UK

Integration of the three genomes in the plant cell: the extent of protein and nucleic acid traffic between nucleus, plastids and mitochondria. F. Quetier[1], A. Brennicke[2] & J. A. Eisen[3].
1, Genoscope, Centre Nationale de Sequencage, 2 rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France; 2, Molekulare Botanik, Universität Ulm, 89069 Ulm, Germany; 3, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Transposable elements. T. Bureau[1], B.-A. Legault[1], Q.-H. Le[1], N. Agrawal[1], Z. Yu[1] & R. Martienssen[2]
1, McGill University, Dept of Biology, 1205 rue Dr Penfield, Montreal, Quebec, H3A 1B1, Canada; 2, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

rDNA, telomeres and centromeres. G. P. Copenhaver[1], S. Luo[1], C. S. Pikaard[2] & D. Preuss[1]
1, Howard Hughes Medical Institute, The University of Chicago, 1103 East 57th Street, Chicago, Illiois, USA; 2, Biology Department, Washington University in St Louis, St Louis, Missouri 63130, USA

Membrane transport. I. T. Paulsen[1] & M. Sussman[2]
1, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA; 2, University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, Wisconsin 53706, USA

DNA repair and recombination. A. B. Britt[1] & J. A. Eisen[2]
1, Section of Plant Biology, University of California, Davis, California 95616, USA; 2, The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA

Gene regulation. D. A. Selinger[1], R. Pandey[1], D. W. Mount[2], V. L. Chandler[1], R. A. Jorgensen[1] & C. Pikaard[3]
1, Department of Plant Sciences, University of Arizona, 303 Forbes Hall; and 2, Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA; 3, Biology Department, Washington University in St Louis, St Louis, Missouri 63130, USA

Cellular organization. G. Juergens
Entwicklungsgenetik, ZMBP-Centre for Plant Molecular Biology, auf der Morgenstelle 1, Tuebingen D-72076, Germany

Development. E. M. Meyerowitz.
Division of Biology, California Institute of Biology, Pasadena, California 91125, USA

Signal transduction. J. R. Ecker[1] & A. Theologis[2].
1, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA; 2, Plant Gene Expression Center/USDA-UC Berkeley, 800 Buchanan Street, Albany, California 94710, USA

Recognition of and response to pathogens. J. Dangl[1] & J. D. G. Jones[2]
1, Biology Department, Coker Hall, University of North Carolina, Chapel Hill, North Carolina 27599, USA; 2, Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UJ, UK

Photomorphogenesis and photosynthesis. M. Chen & J. Chory
Howard Hughes Medical Institute and Plant Biology Laboratory, The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

Metabolism. C. Somerville
Carnegie Institution, 260 Panama Street, Stanford, California 94305, USA

Breakthrough Technologies

Arabidopsis Map-Based Cloning in the Post-Genome Era

Georg Jander*, Susan R. Norris, Steven D. Rounsley, David F. Bush, Irena M. Levin[1], and Robert L. Last

Cereon Genomics LLC, 45 Sidney Street, Cambridge, Massachusetts 02139

Map-based cloning is an iterative approach that identifies the underlying genetic cause of a mutant phenotype. The major strength of this approach is the ability to tap into a nearly unlimited resource of natural and induced genetic variation without prior assumptions or knowledge of specific genes. One begins with an interesting mutant and allows plant biology to reveal what gene or genes are involved. Three major advances in the past 2 years have made map-based cloning in Arabidopsis fairly routine: sequencing of the Arabidopsis genome, the availability of more than 50,000 markers in the Cereon Arabidopsis Polymorphism Collection, and improvements in the methods used for detecting DNA polymorphisms. Here, we describe the Cereon Collection and show how it can be used in a generic approach to mutation mapping in Arabidopsis. We present the map-based cloning of the *VTC2* gene as a specific example of this approach.

Map-based cloning, also called positional cloning, is the process of identifying the genetic basis of a mutant phenotype by looking for linkage to markers whose physical location in the genome is known. The amount of effort required for map-based cloning of genes in Arabidopsis has dropped dramatically in recent years (Fig. 1). Only a few years ago, it was necessary to build a physical map, develop markers, and iteratively zero-in on the gene by "chromosome walking." This was followed by cloning, complementation by transformation, and de novo determination of the sequence of the entire region of interest to high quality without a previously determined wild-type DNA sequence as a guide (Arondel et al., 1992; Giraudat et al., 1992; Leung et al., 1994; Meyer et al., 1994; Mindrinos et al., 1994).

Many of the steps of chromosome walking have been eliminated or have been made much easier by three nearly simultaneous breakthroughs during the past 2 years: sequencing of the entire Columbia (Col-0) Arabidopsis genome (The Arabidopsis Genome Initiative, 2000), the availability of tens of thousands of randomly distributed genetic markers to registered users of the Cereon Arabidopsis Polymorphism Collection (http://www.arabidopsis.org/cereon/), and advances in the methods used to detect DNA polymorphisms. One can now proceed from a mutant with a desirable phenotype to an identified mutation in a gene with less than one person-year of effort (Fig. 1). The minimal start-to-finish time of a mapping project has also been shortened significantly, making it possible to find a gene using an iterative approach taking approximately 1 year (Fig. 2).

In the process of map-based cloning, one starts with a mutant and eventually identifies the gene responsible for the altered phenotype, allowing the plant to tell you what genes are important in the physiological process of interest. This is in contrast to reverse genetic approaches, which tend to rely on some sort of prior knowledge that the gene that is being mutated will be interesting. When using reverse genetic approaches, such as tilling for point mutations (McCallum et al., 2000) or searching for T-DNA insertion mutations (Sussman et al., 2000), one starts with a gene of interest, finds a mutation in that gene, and then looks for a phenotype.

The big advantage to map-based cloning is that it is a process without prior assumptions. Essentially, one is looking at all of the genes in the genome at the same time to find the ones that affect the phenotype of interest. It is a process of discovery that makes it possible to find mutations anywhere in the genome, including intergenic regions and the 40% of Arabidopsis genes that do not resemble any gene with known or inferred function (The Arabidopsis Genome Initiative, 2000).

Insertional mutagenesis using T-DNA or transposons has become increasingly popular as a tool for gene discovery. Pools of lines representing more than 200,000 insertional mutations are available from Arabidopsis stock centers (http://www.Arabidopsis.org/abrc; http://nasc.nott.ac.uk). Large-scale projects are under way for disrupting most genes in Arabidopsis by insertional mutagenesis (Sussman et al., 2000). Mutant screens performed using these populations are undoubtedly worthwhile and can lead to rapid identification of the gene of interest if it is actually has a T-DNA or transposon insertion. However, there are also several good reasons to screen for mutants in chemically mutagenized populations and to isolate the affected genes by map-based cloning.

Insertional mutations tend to result in complete knockouts of the gene, making it difficult to associate a phenotype other than death with essential genes. In contrast, chemical mutagenesis, e.g. with ethyl methane sulfonate, can produce promoter mutations or

[1] Present address: Syrrx Inc., 10410 Science Center Drive, San Diego, CA 92121.

* Corresponding author; e-mail georg.jander@cereon.com; fax 617–551–1950.

www.plantphysiol.org/cgi/doi/10.1104/pp.003533.

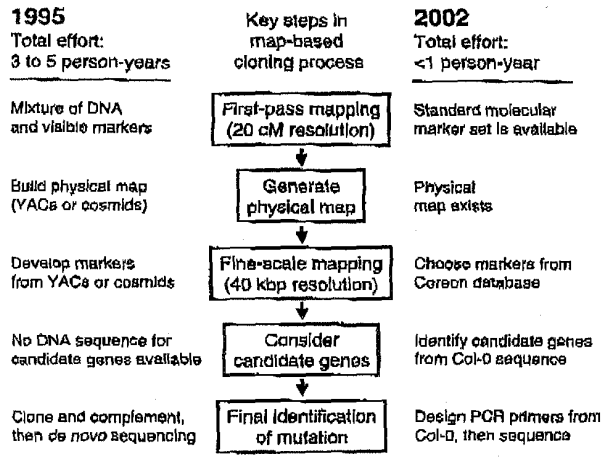

Figure 1. Comparison of effort involved in map-based cloning. Key steps that have become easier between 1995 and 2002 are presented.

mis-sense mutations in the coding region, resulting in a hypomorphic knock-down rather than an amorphic knockout of a protein function. Many interesting but essential genes have been found through such hypomorphic mutations. For instance, "leaky" mutations in *VTC1* (*CYT1*) can result in ozone sensitivity and reduced vitamin C levels in Arabidopsis (Conklin et al., 1999), but knockout mutations cause embryo lethality (Lukowitz et al., 2001). Key regulatory steps in biochemical pathways are often found through dominant point mutations that prevent feedback inhibition of an enzyme, e.g. anthranilate synthase (Kreps et al., 1996; Li and Last, 1996) or Asp kinase (Heremans and Jacobs, 1997). Such dominant mutations would not be found by insertional mutagenesis.

Chemical mutagenesis, in addition to generating a greater diversity of mutations than insertional mutagenesis, also results in many more mutations in each individual plant. Plants mutagenized with T-DNA typically have only one to three insertions per line. Even in a best-case scenario (insertion of three T-DNAs per line in a completely random manner, which is not likely), more than 100,000 plants are needed for a 95% likelihood of having a mutation in a given gene of average size. Screening this many plants can be prohibitive if the mutant screen being performed is laborious or slow. In contrast, ethyl methane sulfonate mutagenesis typically introduces dozens of mutations in each plant line, and it is generally possible to find a mutation in any given gene by screening fewer than 5,000 plants (Feldman et al., 1994).

The techniques of map-based gene identification are also essential for the identification of the genetic basis of phenotypic variation among Arabidopsis ecotypes (natural isolates). The genomes of Arabidopsis ecotypes differ from one another at many thousands of locations and represent a level of genetic variation that is not achievable in the laboratory (Alonso-Blanco and Koornneef, 2000). Hundreds of ecotypes collected from around the world are available to researchers through Arabidopsis stock centers (http://www.Arabidopsis.org/abrc; http://nasc.nott.ac.uk). Phenotypic variation for almost any trait of interest can be found in progeny of crosses made between these ecotypes. In many cases this variation is due to the effects of several genes and is quantitative in nature. Statistical methods developed in the 1990s (Haley and Knott, 1992; Jansen, 1993; Zeng, 1994) and the availability of an almost unlimited set of genetic markers (see below) make it feasible to map and clone such quantitative trait loci (QTL). We will not describe QTL mapping here, but other recent reviews have covered this subject (Kearsey and Farquhar, 1998; Alonso-Blanco and Koornneef, 2000; Yano, 2001).

In this paper, we present a large set of DNA markers identified at Cereon Genomics, we describe how these markers can be applied to a generic map-based cloning project, and we introduce the *VTC2* gene as an example of a specific mapping project.

THE CEREON ARABIDOPSIS POLYMORPHISM COLLECTION

Positional cloning of genes in Arabidopsis is greatly facilitated by the recent sequencing of Col-0 and Landsberg *erecta* (*Ler*). These two ecotypes were sequenced because they are among the most commonly used ecotypes in Arabidopsis research. George Redei, one of the founders of modern Arabidopsis genetics, began working with Col and *Ler* in the 1950s (Redei, 1992). Since then, they have been the subjects of literally thousands of papers that have been published on the genetics, molecular biology, and biochemistry of Arabidopsis. Col-0 and *Ler* are also the parents of a widely used collection of recombinant inbred lines (Lister and Dean, 1993). Hundreds of markers have been analyzed in these lines, and the genetic map produced from this work has become the standard against which other Arabidopsis genetic maps are aligned.

The Col-0 ecotype was the subject of a large international sequencing project, which has produced a nearly complete sequence using a clone by clone approach (The Arabidopsis Genome Initiative, 2000). This high-quality sequence (less than one error in 10,000 bp) is a permanent resource for all future Arabidopsis sequencing efforts. Partial genomic sequence data generated from other ecotypes can be positioned on the framework of Col-0 genome sequence. Sequencing of individual genes from mutants or from other ecotypes has become routine; it is simply a matter of designing PCR primers based on the Col-0 sequence, amplifying the desired gene, and sequencing the product.

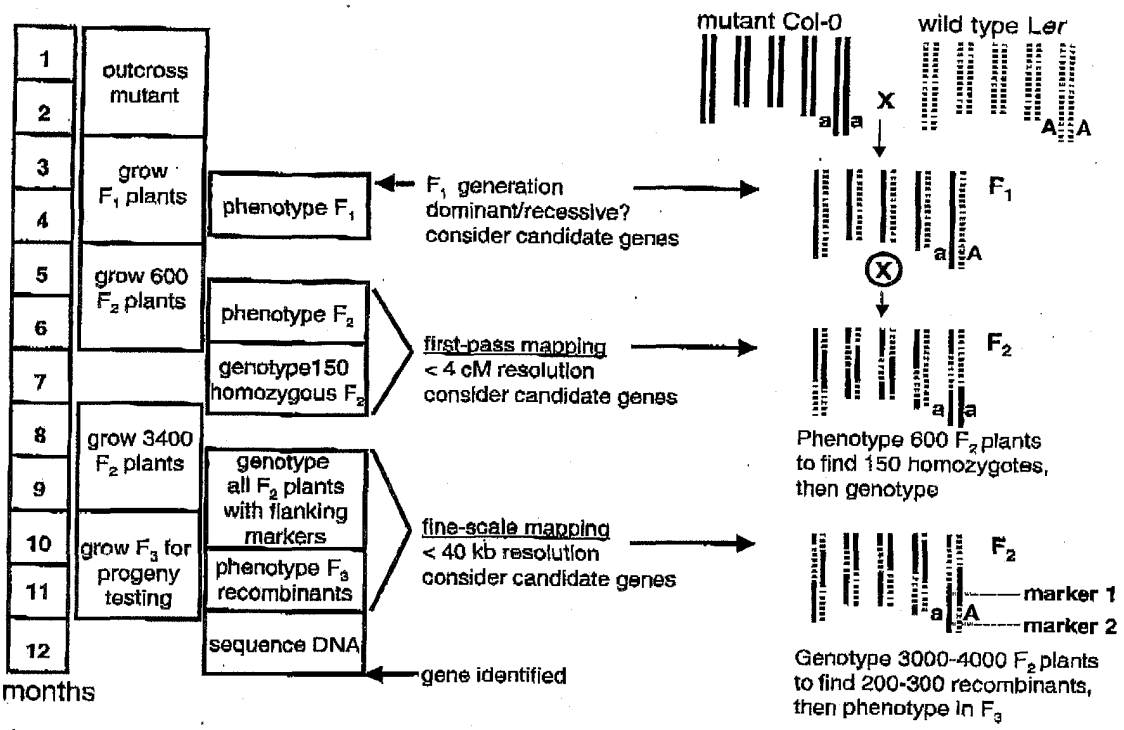

Figure 2. Schematic of the map-based cloning process. Left, Typical 1-year mapping timeline for a mutation whose phenotype can be measured as the plants are growing. Right, Schematic of the five pairs of Arabidopsis chromosomes during critical stages of a sample mapping of a recessive mutation on chromosome 5 in the Col-0 background.

The *Ler* ecotype was the subject of a very different genome sequencing effort, low coverage shotgun sequencing at Cereon Genomics. This project generated approximately 700,000 500-bp sequence traces. Of these, more than 200,000 were chloroplast, mitochondrial, or ribosomal DNA and were not used for the assembly. This left 498,037 traces totaling 263 Mbp of good quality raw sequence, representing approximately 2-fold coverage of the Arabidopsis genome. Assembly of the sequences produced 50,262 contigs (average size, 1.5 kb) and 31,044 single-read sequences. The size of the assembled dataset totaled 92.1 Mbp, suggesting that approximately 70% of the genome is covered at the nucleotide level. To assess the coverage at the gene level, more than 2,000 cDNA sequences from GenBank were extracted and searched against the *Ler* shotgun dataset using the BLASTn algorithm (Altschul et al., 1990). A total of 96.5% of the cDNAs were at least partially detected using a 95% identity cutoff, indicating that at least some sequence from over 95% of all genes is present in the data assembled from the low coverage shotgun approach.

For Arabidopsis researchers who are interested in map-based cloning, the value of two genome sequences greatly exceeds that of only one such sequence. Whereas the availability of the genome sequence of a single ecotype mainly facilitates DNA sequencing in the final stages of a mapping project (Fig. 1), data from two genomes make it possible to develop a database of DNA polymorphisms that can be used as genetic markers. A high-density map of DNA markers greatly facilitates fine-scale genetic mapping. To generate such a map, we compared stretches of *Ler* shotgun sequence with Col-0 genomic sequence determined from cloned bacterial artificial chromosomes (BACs; we will refer to all large DNA clones sequenced by the Col-0 genome project collectively as BACs). Differences between the ecotypes were classified into two types: single nucleotide polymorphism (SNP) changes, which alter a single nucleotide present at specific location in the genome (Fig. 3), and insertion-deletion (InDel) differences, where one ecotype has an insertion of a number of nucleotides relative to the other (Fig. 3).

To detect SNPs and InDels, one must be able to accurately predict true polymorphisms against a background of sequencing errors. This is of particular concern for the *Ler* data, which are unedited shotgun sequence, in contrast to the high quality "finished" Col-0 sequence. To increase the likelihood of detecting real ecotypic differences, fairly stringent criteria were applied to a single base difference before calling it a bioinformatically predicted SNP. The aligned

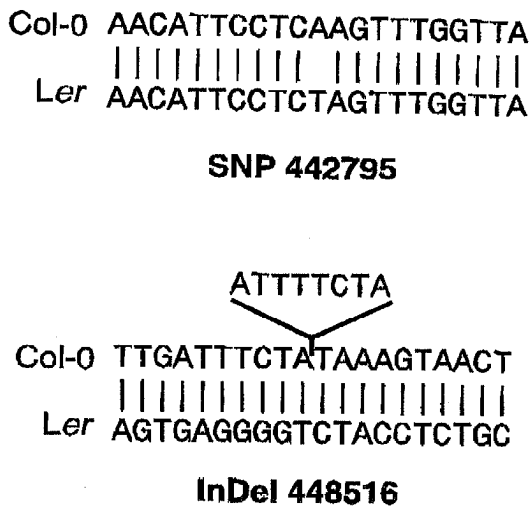

Figure 3. Examples of SNP and InDel polymorphisms. Two markers from the Cereon Arabidopsis Polymorphism Collection are shown. Marker 442795 has a single-nucleotide change from A to T, whereas marker 448516 has an eight-nucleotide insertion in Col-0 versus *Ler*.

region of *Ler* and Col-0 sequence had to be longer than 200 bp and to include more than 75% of the length of the *Ler* sequence. In addition, the polymorphic base must be unambiguous in *Ler*, covered by at least two reads, and be greater than 25 bp from any single coverage region. The quality of the local sequence must be high: The SNP-containing base must have a phrap consensus quality score (Green, 1996, Version 0.980812, downloaded 1999) of at least 40, and the surrounding 25 nucleotides must have consensus scores of at least 30. Re-sequencing of the *Ler* allele of a representative sample of SNPs predicted in this way showed that the success rate was close to 100%. Single-basepair InDels were found using the same methods as those used for SNP prediction. Less stringent criteria were applied for the detection of larger InDels. A gapped alignment between *Ler* and Col-0 was required to be greater than 90% identical over the matched region, with an insertion of at least 2 bp in either Col-0 or *Ler*. Unlike with SNP polymorphisms, we did not confirm a representative sample of predicted InDels by resequencing the *Ler* allele. Given the less stringent selection criteria, the error rate for predicted InDel polymorphisms is likely to be higher than the error rate for predicted SNP polymorphisms.

At the time of writing, sequence for 1,501 Col-0 BACs representing 123 Mbp of Col-0 genome sequence had been compared against the assembled *Ler* shotgun sequence. This resulted in the identification of 37,344 SNPs, 18,579 small InDels (less than or equal to 100 bp), 747 large InDels (larger than 100 bp), or a total of 56,670 polymorphisms. On average, there is one bioinformatically predicted SNP every 3.3 kb and one predicted InDel every 6.6 kb. The

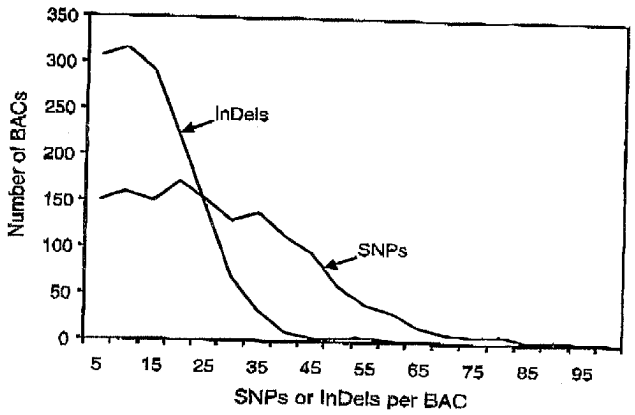

Figure 4. Frequency of SNPs and InDels by BAC. A total of 56,668 SNP and InDel polymorphisms between Col-0 and *Ler* were identified. These polymorphisms were assigned to 1,501 sequenced Col-0 BAC clones (The Arabidopsis Genome Initiative, 2000). Data are presented as bins of 5, i.e. 1 to 5 polymorphisms/BAC, 6 to 10 polymorphisms BAC, etc. Nineteen BACs have no predicted InDel or SNP polymorphisms.

SNPs and InDels are distributed throughout the genome, with most BACs having several polymorphisms that could be used for genetic mapping (Fig. 4). Because of the stringent selection criteria and the partial *Ler* sequence, these numbers represent an underestimate of the true frequency of SNP and InDel differences that exist. For instance, a screen of 500 kb of Arabidopsis sequence by denaturing HPLC (DHPLC) found polymorphisms at a frequency of close to one per kilobasepair (Cho et al., 1999). The Cereon Arabidopsis Polymorphism Collection is made available to registered users at non-profit and educational institutions for non-commercial research. Access is obtained by one-time registration through The Arabidopsis Information Resource Web site (http://www.arabidopsis.org/cereon/). At the time

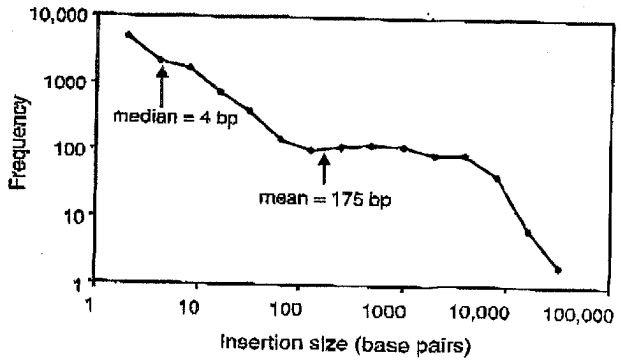

Figure 5. Insertions in Col-0. A total of 10,578 insertions in Col-0 relative to Ler were identified. Insertion size data are presented as bins of 0.3 $\log_{10}$(no. of basepairs), i.e. $\log_{10}$(no. of basepairs) < 0.3, 0.3 < $\log_{10}$(no. of basepairs) < 0.6, etc. The median (4 bp) and mean (175 bp) insertion sizes are indicated.

of writing, 890 researchers from 40 countries had registered to use this database.

The five chromosomes of Arabidopsis have approximately equal densities of SNP polymorphisms. Not surprisingly, SNP frequency varies between exons and introns, with one SNP every 3.1 and 2.2 kb, respectively. Transitions (A/T to G/C) account for 52.8% of the SNPs, and transversions occur with frequencies of 17.4% (A/T to T/A), 23.0% (T/A to G/C), and 7.9% (C/G to G/C). There is no Col-0 or *Ler* bias in the directionality of the transitions or transversions.

InDel polymorphisms between Col-0 and *Ler* range from 1 bp to greater than 38 kb. Due to the average 1.5-kb contig size of the *Ler* shotgun sequence, large insertions can only be detected in the Col-0 background and not in the *Ler* background. Insertions in Col-0 relative to *Ler* have an average size of 175 bp and a median size of 4 bp (Fig. 5). Approximately 10% of the InDels were associated with polymorphisms in the length of simple sequence repeats that were identified with the Sputnik program (Abajian, 1994, downloaded 1999), but most were found in non-repetitive sequences. Most InDels (93%) are smaller than 100 bp, making them suitable for PCR-based marker detection methods (see below).

The Cereon Col-0/*Ler* SNPs and InDels sequences should be very informative for discovering polymorphisms between other ecotype pairs. If one assumes a random genetic reassortment of polymorphisms among Arabidopsis ecotypes, then 50% of the Col-0/*Ler* polymorphisms should be useful for genetic mapping in any other pair of ecotypes. Work done with amplified fragment length polymorphism (AFLP) markers, which generally are due to underlying SNPs, indicates that there is such a random assortment of polymorphisms. Approximately 50% of Col-0/*Ler* AFLP polymorphisms can also be used for segregation analysis in Col-0/C24, Col-0/Wassilewskij, or Col-0/Cape Verde Islands crosses (Peters et al., 2001). Analysis of 79 AFLP markers in 142 ecotypes shows a high degree of recombination in the evolution of these ecotypes, such that it is not possible to draw an "ecotype phylogeny" (Sharbel et al., 2000). Thus, the Cereon Arabidopsis Polymorphism Collection will be useful for mapping QTLs or mutations in most and perhaps all other pairs of Arabidopsis ecotypes. It is a relatively minor disadvantage that one-half of all attempted markers will fail and the average marker density is reduced by 50%, i.e. one SNP every 6.6 kb instead of one SNP every 3.3 kb.

Overall, the density of both SNP and InDel markers is high enough that it is theoretically possible to map most mutations within a few thousand basepairs using either type of marker in any combination of ecotypes. The availability of genetic markers is no longer the limiting factor for the fine-scale genetic mapping needed for map-based cloning in Arabidopsis. Instead, this process is limited by our ability to generate recombination events at a high enough density and to rapidly and inexpensively genotype plants using these markers.

METHODS FOR DETECTION OF DNA POLYMORPHISMS

A critical aspect of map-based cloning is the ability to accurately detect DNA markers at an appropriate cost and throughput. In the past few years, a number of new technologies for high-throughput detection of DNA polymorphisms have been developed. Most of these advances were driven by the field of human genetics, but all of the methods can be applied equally well to plant systems. Because they tend to require a relatively large initial investment, these fast and highly automatable methods are best suited to research settings where large numbers of genotypes need to be determined in a short period of time and with minimal human intervention.

Because SNPs are more common than InDels in biology and are more amenable to automation strategies, most high-throughput genotyping approaches are designed for SNP rather than InDel detection. Oligonucleotide arrays (Gene Chips) contain thousands of oligonucleotides annealed to a glass slide. Such arrays allow the detection of SNP polymorphisms by differential hybridization in a highly parallel and automated manner (Lipshutz et al., 1999). The Taq-Man PCR assay is designed to detect SNPs in a high-throughput manner through the release of fluorescent reporter dye from a quencher on the same oligonucleotide by 5′ nuclease activity (Livak, 1999). By using more than one reporter dye, it is possible to detect different alleles of a SNP in a single reaction. The relatively high price of oligonucleotides tagged with reporter and quencher dyes makes this method cost-effective only if a large number of reactions need to be run with each SNP marker. In pyrosequencing, an enzymatic cascade and luminometric detection system is used to measure the pyrophosphate that is released as a result of nucleotide incorporation (Ahmadian et al., 2000; Alderborn et al., 2000). Because 20 or more nucleotides are determined by this method, it is possible to detect several closely linked SNPs at once. The pyrosequencing method can be automated but has the disadvantage that it does not work well on stretches of repeated nucleotides. DHPLC allows the detection of SNPs through different retention time of heteroduplex and homoduplex DNA in reversed-phase HPLC under partially denaturing conditions (Spiegelman et al., 2000). DHPLC allows detection of SNP polymorphisms in PCR-amplified DNA up to about 1,000 bp in size. Although not inherently high-throughput, DHPLC lends itself nicely to bulked segregant analysis. The method of fluorescence resonance energy transfer combines PCR and oligonucleotide ligation to detect SNPs (Chen et al., 1998). Dye-labeled oligonucleotide

probes are used in this assay, and allele-specific ligation is detected by fluorescence resonance energy transfer, which only occurs when two dye-labeled oligos are joined by ligation. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry can be used to rapidly detect SNPs in short DNA pieces by differences in molecular mass (Wada and Yamamoto, 1997).

A disadvantage of most high-throughput methods for detecting DNA polymorphisms is the high initial equipment cost, which results in a high per-assay cost for a lab that does not need to perform large numbers of genotyping reactions on a routine basis. In contrast, both InDels and SNPs can be detected using gel-based methods, which have a relatively low start-up cost and moderate throughput. InDels of small to moderate size can be detected by PCR amplification and gel electrophoretic separation (Bell and Ecker, 1994). Pairs of PCR primers are designed to amplify a segment of DNA spanning the InDel, and size differences in the amplified products are detected using either agarose or acrylamide gels. Agarose gels are easier and less expensive to use, but size differences of less than 5 bp are difficult to detect reliably. Acrylamide gels, on the other hand, give single-basepair resolution and allow the detection of even very small InDels. In either case, InDels are scored as codominant markers with one band seen on the gel for either homozygous class and two bands seen for heterozygous individuals. To reduce the number of PCR reactions needed for a mapping project, it is possible to pool DNA samples for bulked segregant analysis (Michelmore et al., 1991; Lukowitz et al., 2000), or multiple primer pairs can be added to one reaction tube to amplify several markers at once (Ponce et al., 1999).

Several gel electrophoresis-based strategies for detecting SNP markers have been devised. Many SNPs alter sites cleaved by restriction enzymes and can be used as cleaved-amplified polymorphic sequence (CAPS; Konieczny and Ausubel, 1993) markers. CAPS markers are amplified by PCR, the amplified DNA is cleaved with the appropriate restriction enzyme, and the cleavage products are examined on agarose gels. Just as with InDels, such markers are codominant, allowing the differentiation of heterozygotes and either homozygote class. If there is no suitable restriction site at a SNP, it is possible create a site during PCR amplification with suitably designed primers (dCAPS [Michaels and Amasino, 1998; Neff et al., 1998]). Disadvantages of using CAPS and dCAPS for genotyping include the extra time and cost involved in the restriction enzyme digestion and the possibility of a false result attributable to incomplete digestion by the restriction enzyme.

It is also possible to detect SNPs using allele-specific PCR primers, where the 3′ end of a primer has a perfect match with one allele and a mismatch with the other allele (Ugozzoli and Wallace, 1991). In theory, such primers can be used to preferentially amplify one allele of a SNP, but in practice a single-basepair change is often not enough to allow reliable differentiation between the two alleles of an SNP (Kwok et al., 1990; Cha et al., 1992). A modification of the allele-specific amplification procedure (single nucleotide amplified polymorphism [SNAP]) has recently been described (Drenkard et al., 2000). In this method, additional mismatches are introduced in the amplifying primers to maximize the difference in the amplification efficiencies of the two alleles of the SNP. Primer basepair changes that allow differential amplification of SNP sites can be predicted using the SNAPER program. Both the SNAPER program and a collection of primers that have been used successfully to amplify Arabidopsis SNAP markers can be found at http://patho.mgh.harvard.edu/ausubel-web. As with CAPS, SNAP markers are codominant and can be detected on agarose gels. However, it is necessary to run two PCR reactions—one for each allele of the SNP—to get complete SNAP genotyping data.

The detection of SNPs and InDels is an essential part of the map-based cloning process. Because marker discovery is no longer a problem in Arabidopsis, the selection of an efficient genotyping platform plays a critical role in the mapping timeline that we describe in the next section. We have mentioned several commonly used genotyping methods, and the choice of which method to apply will depend on the resources of an individual laboratory and the number of genotyping reactions that will need to be performed.

MAP-BASED CLONING PROCESS

Given a sequenced genome and a dense collection of genetic markers, map-based cloning becomes a relatively straightforward process. Figure 2 illustrates a time-efficient approach to map-based cloning in Arabidopsis, a variant of the "chromosome landing" method proposed by Tanksley et al. (1995). Starting with a mutation in the Col-0 or *Ler* background, it is possible to proceed from having a mutant plant to identifying the affected gene in approximately 1 year. The overall length of this cloning process is dictated largely by the fact that it incorporates five cycles of plant growth (we assume 2 months/cycle).

As a first step in the mapping process, the mutant is out-crossed to the opposite ecotype (Col-0 or *Ler*). In most cases, it is not necessary to "clean up" the genetic background of the mutant by back-crossing and it does not matter whether the mutant is used as the male or the female parent in the out-cross. F_1 seeds are planted and, as the plants are growing, it is possible to perform phenotype and genotype analysis. Presence or absence of the phenotype in the F_1 generation will suggest whether the mutation of in-

terest is likely to be dominant or recessive. We recommend genotyping the F_1 plants with a few markers to make sure that they are truly heterozygous and that there was no mistake made during the cross. Similarly, it is worthwhile to genotype the original mutant to make sure that it is in the presumed ecotype background. Contamination with other ecotypes is a surprisingly frequent cause of "mutants" that arise in screens.

F_2 seeds are collected from self-pollination of the F_1 plants, and a population of approximately 600 individuals is planted for first-pass mapping (Fig. 2). As they are growing, the phenotype of the F_2 plants is determined, unless the trait can only be scored in the progeny (F_3) seed. It should be possible to identify approximately 150 plants in this population as homozygous: homozygous mutant in the case of a recessive mutation or homozygous wild type in the case of a dominant mutation. DNA for genotype analysis is prepared from the leaves or other tissue of these 150 plants. Initially, the 150 plants are genotyped with 25 markers, spaced roughly every 20 centiMorgan (cM) apart on the five chromosomes. Genetic linkage to one or more of the 25 markers is determined and a three-point cross is used to define a 20-cM interval that contains the gene of interest. Once a 20-cM interval has been found, additional markers are used to narrow down the region of interest to approximately 4 cM. Given a population of 150 plants, it should be possible to determine this 4-cM interval with a high degree of certainty. The two markers closest to the mutation on either side will be used as flanking markers in further work.

Next, it is necessary to plant a larger F_2 population for fine-resolution mapping (Fig. 2). The ultimate goal of fine mapping is to narrow down the region containing the gene of interest to 40 kb or less (approximately 0.16 cM genetic distance in Arabidopsis). There would ideally be several recombination events in this interval to define the position of the mutation that is being mapped. Unfortunately, the number of F_2 plants needed to have a 95% chance of recombination events in a given genetic interval increases rapidly as the size of the interval decreases (Fig. 6). We recommend having a fine mapping population of 3,000 to 4,000 plants (including the original 600 lines grown for first-pass mapping) to give a high probability of mapping the gene of interest to less than 40 kb. In areas of the genome with reduced meiotic recombination, e.g. near the centromeres, larger F_2 populations will be necessary to map a mutation to an equivalent physical interval on the chromosome. Many Arabidopsis mapping projects have been successful with fewer than 3,000 to 4,000 F_2 plants (Lukowitz et al., 2000), but when planting fewer plants one runs the risk of extending the mapping timeline by having to plant an additional F_2 population later on.

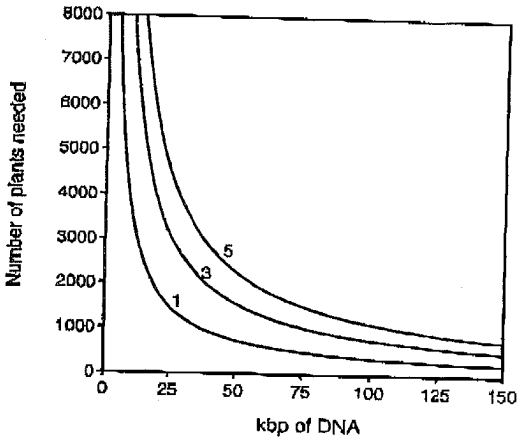

Figure 6. Number of plants needed to find recombinants. The curves show the number of F_2 plants needed to have a 95% chance of finding at least one plant (1), at least three plants (3), or at least five plants (5) with recombination events in a given physical interval of DNA. The calculations assume an average 250 kb/cM for Arabidopsis (Lukowitz et al., 2000). The possibility of multiple recombination events in one individual plant has a negligible effect and is not included in the calculations.

At this point, plants that are recombinant in the 4-cM interval determined by first-pass mapping are sought for use in fine mapping. DNA is isolated from the mapping population of 3,000 to 4,000 plants and the genotype of the two flanking markers is determined. This should identify 200 to 300 plants that have genetic recombination events in the region of interest (Fig. 2). The allelic state of the mutation being mapped (homozygous mutant, homozygous wild type, or heterozygous) in these recombinant plants is determined by looking at the phenotype in a representative sample of progeny in the F_3 generation. Additional markers in the 4-cM interval are used to look for increasingly tight linkage to the mutation. In most cases, it should be possible to define a pair of markers flanking the mutation that are less than 40 kb apart.

Once an interval of less than 40 kb containing the mutation of interest has been determined, this entire region is sequenced to find the mutation. In theory, it is possible to map a mutation to the single-gene level using the Cereon Arabidopsis Polymorphism Collection, but the number of F_2 plants needed to find recombinants in such a small interval would be very large (Fig. 6). It is faster and less expensive to sequence a larger interval. Because the sequence of the Col-0 genome is known, one efficient way to sequence the mutant region is to design PCR primers to amplify overlapping segments of about 500 bp spanning the entire 40 kb. These segments are then sequenced and assembled, the sequence is compared with that of a wild-type plant (Col-0 or *Ler*), and the mutation is identified. In the case of a mutation in the *Ler* background, it is necessary to also sequence

the *Ler* wild type for comparison at every location where a difference to the wild-type Col-0 is found. In the case of a mutation in Col-0, a published sequence is available. However, it is necessary to confirm that any nucleotide that diverges from the published Col-0 sequence was induced by the mutagenesis treatment and is not present in the wild-type progenitor strain. This is because strain to strain differences exist in "Col-0 wild type," and even at the high quality standard of the Col-0 sequence, sequencing errors are expected and found.

APPLICATION OF CEREON MARKERS TO CLONING *VTC2*

The identification of the *VTC2* gene is a specific example of a map-based cloning project using the Cereon Arabidopsis Polymorphism Collection. The *vtc2-1* mutation was isolated in a screen for ozone-sensitive mutants of Arabidopsis (Conklin et al., 1996). Further work showed that this mutant was deficient in ascorbic acid (vitamin C), and an additional three alleles (*vtc2-2*, *vtc2-3*, and *vtc2-4*) were isolated based on this phenotype. A first-pass map position for the *vtc2-1* mutation between CAPS markers WU95 (74 cM) and PRHA (78 cM) on chromosome 4 was reported (Conklin et al., 2000).

The CAPS markers WU95 and PRHA are relatively difficult to score. Instead, we used the nearby InDel markers (449235 and 450577 from the Cereon Arabidopsis Polymorphism Collection) as flanking markers for fine mapping (Fig. 7A). These markers are approximately 980 kb apart on chromosome 4. DNA segments spanning these markers were amplified by PCR, and the amplified products were detected by PAGE. A population of 3,700 Col-0 *vtc2-2* × *Ler* F_2 plants was analyzed with markers 449235 and 450577. A total of 52 recombinants were identified and confirmed by repeating the genotyping with the same markers in the F_3 generation. The number of recombinants is considerably less than one would expect given the genetic separation previously reported for the CAPS markers WU95 and PRHA (4 cM apart, expected approximately 280 recombinants). We do not have a good explanation for this observation, but it does illustrate the utility of generating a mapping population that is larger than the theoretical minimum needed.

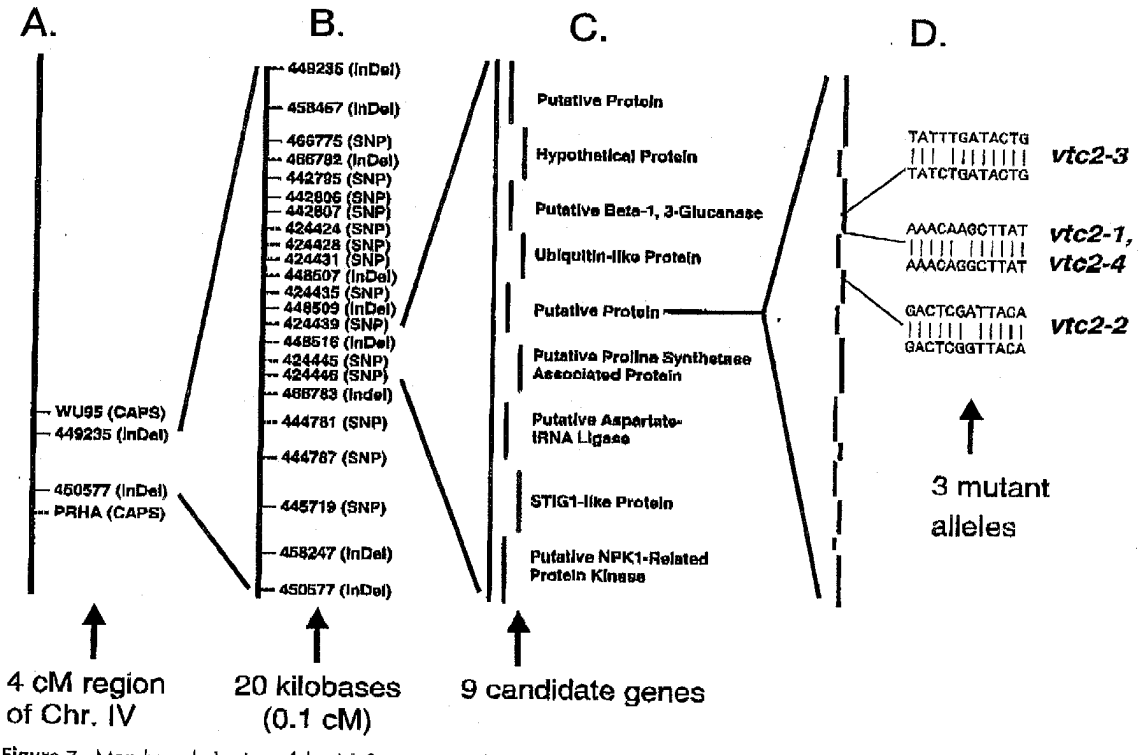

Figure 7. Map-based cloning of the *VTC2* gene. A, First-pass mapping of the *VTC2* identified flanking CAPS markers WU95 and PRHA (4 cM apart). B, Fine mapping of *VTC2* using SNP and InDel markers identified markers 424439 and 424446 in the Cereon Arabidopsis Polymorphism Collection (20 kb apart) as the closest flanking markers based on the available recombinants. C, Nine candidate genes between the SNP markers 424439 and 424446 were identified from the Col-0 sequence in GenBank. D, Mutations in *vtc2-1*, *vtc2-2*, *vtc2-3*, and *vtc2-4* were identified by sequencing. Staggered lines represent the predicted exons and introns of the *VTC2* gene. The 5′ end of the gene is at the bottom.

Additional markers between 449235 and 450577 were chosen from the Cereon Arabidopsis Polymorphism Collection (Fig. 7B) for fine mapping. All 52 recombinants were genotyped with these 21 markers to narrow down the positions of the recombination events. Pieces of DNA containing the marker of interest were amplified by PCR, and the polymorphisms were detected by PAGE (for InDels) or DNA sequencing (for SNPs). Vitamin C levels of individual F_3 progeny (at least 20 per line) were measured to determine whether the 52 F_2 recombinants were homozygous mutant, homozygous wild type, or heterozygous at the *VTC2* locus. This information was combined with the marker data to identify markers 424439 and 424446, which are contained in BAC F10 M23 (GI:4756963), as the closest markers flanking the mutation.

Markers 424439 and 424446 are approximately 20 kb apart. In the Col-0 genomic sequence, there are nine predicted genes in this region (Fig. 7C), but none are annotated as enzymes of the proposed Wheeler-Smirnoff Pathway for vitamin C biosynthesis in plants (Wheeler et al., 1998). We designed primer pairs to amplify overlapping segments of DNA spanning the 20-kb region from the *vtc2-2* mutant. Sequencing of these fragments and comparison with the wild-type Col-0 sequence identified a mis-sense change in the putative gene F10M23.190 (GI:7452423; Fig. 7D), resulting in a Gly to Asp change in the predicted exon 5 (new GenBank ID AF508793). This gene was also sequenced from the three other *vtc2* mutants. A mis-sense mutation was identified in *vtc2-3* (Fig. 7D), resulting in a Ser to Phe change in the predicted exon 6. Both *vtc2-1* and *vtc2-4* had the same mutation, which changed the 3′ splice site of the predicted intron 5 from AG to AA (Fig. 7D). These last two mutations are almost certainly independently generated, because one was isolated in wild-type Col-0 and a the other was from a strain of Col-0 carrying a *PAT1-GUS* transgene (Rose and Last, 1997). Together, these four mutations show that putative gene F10 M23.190 is *VTC2*. As additional confirmation, all four mutant alleles of *VTC2* were complemented using genomic clones of F10M23.190 isolated from Col-0 by PCR (I. Levin and S. Norris, unpublished data).

The F10M23.190 gene (*VTC2*) was previously annotated as an undefined protein (GI:7452423; Mayer, 1999). The most similar proteins in the GenBank database are as follows: Arabidopsis protein MCO15.7, *Caenorhabitis elegans* protein C10F3.4, and fruitfly (*Drosophila melanogaster*) protein CG3552, none of which have a demonstrated function. Thus, although we have a phenotype associated with mutations in *VTC2*, the regulatory or biosynthetic pathways leading to the reduced vitamin C levels in these mutants remain to be discovered.

DISCUSSION

We have outlined a map-based cloning strategy, which leads to the identification of an Arabidopsis mutation in a straightforward manner in approximately 1 year. Our timeline assumes that it is possible to determine the phenotype of F_2 plants as they are growing. If the phenotype of interest is measured on seeds (i.e. F_3 seeds from F_2 plants), then the mapping time will be increased by 3 months. The strategy that we propose is designed to minimize the number of plants that have to be subjected to phenotypic analysis. In most cases, DNA based markers can be determined faster and more accurately than individual plant phenotypes. Obviously, if phenotyping is easier than genotyping, this procedure can be changed by identifying a large number of homozygous mutant, or wild type in the case of dominant mutations, plants and genotyping these alone.

Modifications of the process that we have outlined can speed up the mapping timeline. In many cases, as the mapping region is narrowed down, candidate genes become obvious, and it is possible to shift to sequencing at any stage during the process (Fig. 2). For rare examples of very reliable phenotypes, it may not be necessary to grow an F_3 generation for progeny testing, thus, shortening the timeline by 2 months. It is also possible to grow a single large F_2 population, rather than two sequentially grown populations (first-pass mapping and fine-scale mapping). However, this may result in wasted effort because some mutations are recalcitrant to genetic mapping. Situations that can make a given mutation difficult or impossible to map include: QTL variation for the trait of interest in the Col-0/*Ler* F_2 population, phenotypes caused by multiple mutations, sensitivity of the phenotype to environmental variation in the greenhouse or growth chamber, and non-nuclear mutations.

The mapping timeline that we have outlined depends on the ability to rapidly genotype large numbers of plants. It may be difficult to maintain this timeline by using gel-based methods for SNP and InDel detection. High-throughput SNP detection methods are available, but they involve a high initial equipment cost that could make them prohibitive to set up and use in an individual laboratory. One solution to this problem may be for universities or academic departments to set up genotyping centers, similar to those that currently exist for DNA sequencing. Similar to a DNA sequencing center, a genotyping center could serve a large number of researchers working in all areas of molecular genetics.

The current rate-limiting step for map-based cloning in Arabidopsis is the number of F_2 plants that must be analyzed to find recombinants in a sufficiently small interval of DNA. There are no known methods for increasing meiotic recombination frequency in Arabidopsis (or any other plant). However, both ecotype-specific variation (Barth et al.,

2001) and mutations that decrease meiotic recombination frequency (Masson and Paszkowski, 1997; Grelon et al., 2001) have been reported. It is plausible that it will be possible to selectively alter meiotic recombination frequency at some point in the not too distant future by crossing QTLs from other ecotypes into standard laboratory strains, by overexpressing proteins necessary for elevated meiotic recombination, or perhaps by physical or chemical treatments that increase the recombination rate.

Sequencing of the Arabidopsis genome, the availability of the Cereon Arabidopsis Polymorphism Collection, and advances in the methods used for DNA marker detection have made map-based cloning of mutations in Arabidopsis a routine process. Mutation mapping will play a central role in the process of assigning a function to the thousands of plant genes that currently are known only as predicted open reading frames. Given the advantages of map-based cloning that we have outlined in the introduction, this is a viable approach for gene discovery that can be used in any laboratory.

Received January 31, 2002; accepted February 5, 2002.

LITERATURE CITED

Abajian C (1994) Sputnik program, http://rast.abajian.com/sputnik/

Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyren P, Uhlen M, Lundeberg J (2000) Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem **280:** 103–110

Alderborn A, Kristofferson A, Hammerling U (2000) Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res **10:** 1249–1258

Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. Trends Plant Sci **5:** 22–29

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Arondel V, Lemieux B, Hwang I, Gibson S, Goodman HM, Somerville CR (1992) Map-based cloning of a gene controlling omega-3 fatty acid desaturation in Arabidopsis. Science **258:** 1353–1355

Barth S, Melchinger AE, Devezi-Savula B, Lubberstedt T (2001) Influence of genetic background and heterozygosity on meiotic recombination in *Arabidopsis thaliana*. Genome **44:** 971–978

Bell CJ, Ecker JR (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics **19:** 137–144

Cha RS, Zarbl H, Keohavong P, Thilly WG (1992) Mismatch amplification mutation assay (MAMA): application to the c-H-ras gene. PCR Methods Appl **2:** 14–20

Chen X, Livak KJ, Kwok PY (1998) A homogeneous, ligase-mediated DNA diagnostic test. Genome Res **8:** 549–556

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N et al. (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet **23:** 203–207

Conklin PL, Norris SR, Wheeler GL, Williams EH, Smirnoff N, Last RL (1999) Genetic evidence for the role of GDP-mannose in plant ascorbic acid (vitamin C) biosynthesis. Proc Natl Acad Sci USA **96:** 4198–4203

Conklin PL, Saracco SA, Norris SR, Last RL (2000) Identification of ascorbic acid-deficient *Arabidopsis thaliana* mutants. Genetics **154:** 847–856

Conklin PL, Williams E, Last RL (1996) Environmental stress sensitivity of an ascorbic acid-deficient Arabidopsis mutant. Proc Natl Acad Sci USA **93:** 9970–9974

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiol **124:** 1483–1492

Feldman KA, Malmberg RL, Dean C (1994) Mutagenesis in Arabidopsis. *In* E Meyerowitz, C Somerville, eds, Arabidopsis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 137–182

Giraudat J, Hauge BM, Valon C, Smalle J, Parcy F, Goodman HM (1992) Isolation of the Arabidopsis *ABI3* gene by positional cloning. Plant Cell **4:** 1251–1261

Green P (1996) Phrap program, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html

Grelon M, Vezon D, Gendrot G, Pelletier G (2001) *AtSPO11-1* is necessary for efficient meiotic recombination in plants. EMBO J **20:** 589–600

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324

Heremans B, Jacobs M (1997) A mutant of *Arabidopsis thaliana* (L.) Heynh. with modified control of aspartate kinase by threonine. Biochem Genet **35:** 139–153

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211

Kearsey MJ, Farquhar AG (1998) QTL analysis in plants; where are we now? Heredity **80:** 137–142

Konieczny A, Ausubel F (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J **4:** 403–410

Kreps JA, Ponappa T, Dong W, Town CD (1996) Molecular basis of alpha-methyltryptophan resistance in *amt-1*, a mutant of *Arabidopsis thaliana* with altered tryptophan metabolism. Plant Physiol **110:** 1159–1165

Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C, Sninsky JJ (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res **18:** 999–1005

Leung J, Bouvier-Durand M, Morris PC, Guerrier D, Chefdor F, Giraudat J (1994) Arabidopsis ABA response

gene *ABI1*: features of a calcium-modulated protein phosphatase. Science **264:** 1448–1452

Li J, Last RL (1996) The *Arabidopsis thaliana trp5* mutant has a feedback-resistant anthranilate synthase and elevated soluble tryptophan. Plant Physiol **110:** 51–59

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet Suppl **1:** 20–24

Lister C, Dean C (1993) Recombinant inbred line for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J **4:** 745–750

Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal **14:** 143–149

Lukowitz W, Gillmor CS, Scheible WR (2000) Positional cloning in Arabidopsis: why it feels good to have a genome initiative working for you. Plant Physiol **123:** 795–805

Lukowitz W, Nickle TC, Meinke DW, Last RL, Conklin PL, Somerville CR (2001) Arabidopsis *cyt1* mutants are deficient in a mannose-1-phosphate guanylyltransferase and point to a requirement of N-linked glycosylation for cellulose biosynthesis. Proc Natl Acad Sci USA **98:** 2262–2267

Masson JE, Paszkowski J (1997) *Arabidopsis thaliana* mutants altered in homologous recombination. Proc Natl Acad Sci USA **94:** 11731–11735

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402:** 769–777

McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeted screening for induced mutations. Nat Biotechnol **18:** 455–457

Meyer K, Leube MP, Grill E (1994) A protein phosphatase 2C involved in ABA signal transduction in *Arabidopsis thaliana*. Science **264:** 1452–1455

Michaels SD, Amasino RM (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J **14:** 381–385

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA **88:** 9828–9832

Mindrinos M, Katagiri F, Yu GL, Ausubel FM (1994) The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78:** 1089–1099

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant J **14:** 387–392

Peters JL, Constandt H, Neyt P, Cnops G, Zethof J, Zabeau M, Gerats T (2001) A physical amplified fragment-length polymorphism map of Arabidopsis. Plant Physiol **127:** 1579–1589

Ponce MR, Robles P, Micol JL (1999) High-throughput genetic mapping in *Arabidopsis thaliana*. Mol Gen Genet **261:** 408–415

Redei GP (1992) A heuristic glance at the past of Arabidopsis genetics. *In* C Koncz, N Chua, J Schell, eds, Methods in Arabidopsis Research. World Scientific, Singapore, pp 1–15

Rose AB, Last RL (1997) Introns act post-transcriptionally to increase expression of the *Arabidopsis thaliana* tryptophan pathway gene *PAT1*. Plant J **11:** 455–464

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol **19:** 2109–2118

Spiegelman JI, Mindrinos MN, Oefner PJ (2000) High-accuracy DNA sequence variation screening by DHPLC. Biotechniques **29:** 1084–1092

Sussman MR, Amasino RM, Young JC, Krysan P, Austin-Phillips S (2000) The Arabidopsis knockout facility at the University of Wisconsin-Madison. Plant Physiol **124:** 1465–1467

Tanksley SD, Ganal MW, Martin GB (1995) Chromosome landing: a paradigm for map-based cloning in plants with large genomes. Trends Genet **11:** 63–68

Ugozzoli L, Wallace RB (1991) Allele-specific polymerase chain reaction. Methods Enzymol **2:** 42–48

Wada Y, Yamamoto M (1997) Detection of single-nucleotide mutations including substitutions and deletions by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom **11:** 1657–1660

Wheeler GL, Jones MA, Smirnoff N (1998) The biosynthetic pathway of vitamin C in higher plants. Nature **393:** 365–369

Yano M (2001) Genetic and molecular dissection of naturally occurring variation. Curr Opin Plant Biol **4:** 130–135

Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468

MONSANTO

#11 CD 11/12/02

Mystic Research
Monsanto Company
62 Maritime Drive, Mystic, CT 06355-1958
Telephone: 860-572-5200 Fax: 860-572-5280

DATE: November 8, 2002

TO: Examiner Juliet Caroline Einsmann

COMPANY: USPTO Art Unit 1634

FAX NUMBER: 703 308-4242

FROM: Donna Scherer

cc:

Total number of pages to transmit, including cover page: 45

If you do not receive this fax legibly, call the receptionist at 860-572-5200.

Re: USSN 09/803,736
Filed: 3/12/2001

Restriction Response

PTO/SB/97 (08-00)
Approved for use through 10/31/2002. OMB 0851-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

38-10(15493)D

Certificate of Transmission under 37 CFR 1.8

I hereby certify that this correspondence is being facsimile transmitted to the United States Patent and Trademark Office

on November 8, 2002
Date

Donna E. Scherer
Signature

Donna E. Scherer
Typed or printed name of person signing Certificate

Note: Each paper must have its own certificate of transmission, or this certificate must identify each submitted paper.

In re Application of : David F. Bush et al.
Serial No. 09/803,736
Filed: March 12, 2001
For: PLANT POLYMORPHIC MARKERS AND USES THEREOF

Papers submitted:
RESPONSE TO RESTRICTION REQUIREMENT AND REQUIREMENT FOR INFORMATION UNDER 37 CFR 1.105 (6 pages)
Publications:
Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature 408:796-814 (December 14, 2000) (20 pages)
Jander et al. (2002) Arabidopsis Map-Based Cloning in the Post-Genome Era, Plant Physiology 129:440-450 (11 pages)
Printout from www.arabidopsis.org/cereon Arabidopsis marker website (3 pages)
Excerpts from Table A with information on markers 466799 and 474736 (3 pages)